

[Intro slide.] Good evening. I am Eliezer Yudkowsky, cofounder and current Research Fellow of the Singularity Institute for Artificial Intelligence, a 501(c)(3) nonprofit located in Silicon Valley. I'd like to talk to you today about the most powerful force in the known universe.

[Brain] In our skulls we each carry 3 pounds of slimy wet grey stuff, corrugated like crumpled paper. If you didn't know anything about human anatomy, if you didn't know what a brain was, and you saw it lying in the street, you'd probably say "Yuck" and try not to get it on your shoes. Aristotle thought the brain was an organ that cooled the blood. It doesn't *look* impressive. It doesn't look big [skyscraper], or dangerous [sword], or beautiful [necklace]. It does look like it might be complicated [brain], but nowhere near as complicated as when you look inside. [van Essen] This, for example, is a diagram of the primate visual system. This is the data flow between the major modules in the software that lets you look around and see things.

[Effects.] A skyscraper, a sword, a crown, a gun, a nuclear weapon, a space shuttle, a dollar bill, a computer, all these are byproducts of the wet grey thing. They popped out of the brain like a jack from a jack-in-the-box. Almost everything you see in the room around you, the chairs, your clothes, the ceiling, are effects caused by human intelligence. Point to something in your environment and ask "Why is it that way?" and, often enough, the answer is "intelligence". Human beings leave behind patterns like smoke puffs from an engine. We breathe out complexity and sweat order. There's a very powerful trick inside that grey lump, a very powerful trick embedded in all those complicated neural circuits. A space shuttle is an impressive trick, a nuclear weapon is an impressive trick, an automated factory is an impressive trick - but none are as impressive as the master trick, the brain trick, the trick that does all these other tricks at the same time.

In everyday life, we take our human intelligence for granted because everyone has it. I mean, suppose everyone had the powers of Spiderman. [Spiderman.] Suppose we could all shoot webs from our hands and climb buildings. Then no one would notice; no one would think it was important. We all have a superpower that's much more impressive than webshooting, but, since we all have the same superpower, we forget how powerful it is. People say things like "intelligence is no match for a gun", as if guns had grown on trees.

[Book smarts.] In everyday life, when you say the word "intelligence", people think of book smarts - calculus, chess, memorizing lists of facts, applying strict rules to well-understood situations. We imagine the starving professor with an IQ of 160 and the billionaire CEO with an IQ of merely 120. It takes more than book smarts to succeed in the human world: Persuasiveness, enthusiasm, empathy, strategic thinking, musical talent, rationality, thinking on your feet - but note that every factor I just listed is cognitive. Empathy happens in the brain, not the kidneys. There are not many famous novelists, or famous military generals, or

famous politicians who are monkeys. Intelligence is the foundation of human power, the strength that fuels our other arts. If you don't have the human superpower you're not even in the game.

[Parochial.] People think about intelligence as if it were a scale that ran from village idiot to Einstein. We forget the things we have in common, and see only our differences. IQ tests, for example, are things that you can only administer to humans. A mouse would just eat the IQ test. If you can take an IQ test for humans, at all, in the first place, you've already established yourself as one of the smartest beings on the face of the Earth. This scale from village idiot to Einstein is measuring centimeter height differences within a species of giants. When you hear the word "intelligence", don't think of Einstein, think of humans.

[Cosmopolitan.] When I talk about "intelligence", I'm talking about the scale that runs from rocks to humans. On that scale, the whole human species is packed into a small dot. Human beings, in general, have three times as much brain and six times as much frontal cortex as a primate our size. That's true whether you're a village idiot or Einstein; there's variance, but not that much variance.

Let us turn now to matters of transhumanism. To paraphrase the famous science fiction author Robert Heinlein [Heinlein], a transhuman is a transhuman mind; anything else is a distraction. If you can give people the ability to control an extra pair of arms, that's really cool, and you *have* benefited humanity, but you haven't created transhumans. If you invent artificial red blood cells that let people hold their breath for four hours, you can save many people who now die from strokes and heart attacks. That is a true and worthy endeavor. But it does not create a transhuman. Humanity did not rise to prominence upon Earth by holding its breath longer than other species. If you can give someone the ability to shoot webs from their hands - why even bother? All you'd do is create a mere superhero.

In today's society, most people get their idea of what's important from advertising. Madison Avenue would have you believe that *important* technologies [iPhone.] are new, improved, shiny, with extra features, and expensive. They want to sell you products by calling them "futuristic", and so people get the idea that the future is about new products. Hollywood movies tell you you're in the Future by showing you soaring skyscrapers and glittering gadgets. In movies, the Future is like Oz, a distant land full of curious people and curious customs. We all know what the Future looks like [Future], even though we've never been there, the same way we know what Oz looks like. And the futurists who make their living as entertainers, who tell fantastic stories to an audience, are part of the same Hollywood, Madison Avenue culture - they make comfortable prophecies that in the future you will be able to buy ever more cool devices, maybe with a side order of improved health care.

And this culture treats intelligence enhancement the same way, like an upgrade you could buy for your iPhone. In movies, if anyone has a direct brain interface to a computer, that's just part of the background scenery that tells you you're in the Future, like people wearing strange clothes. [BCI.] You'll be able to buy brain implants just like you can buy botox injections, and you'll use your brain implant to play expensive video games.

[Effects.] But intelligence is not just another commodity. Intelligence is the root of all our technology. All our impressive gadgets grow out of that. When you mess with intelligence you shake the technology tree by its roots. If drugs or neurosurgery or gene therapy can make you smarter, that is not just another bit of gee-whiz sparkly technology. It is messing with the master trick, the brain trick, the first cause of which all technology is an effect. [Unlike other.] Artificial Intelligence is not like interplanetary travel, a cancer cure, or nanomanufacturing. Progress in brain-computer interfaces does not belong on the same graph that shows faster aircraft, smaller MP3 players, or better medical care. One of these technologies is not like the others, one of these technologies doesn't belong.

If you want to understand the future, look to the cognitive technologies. Look to the technologies that impact upon the mind.

A common reaction to this notion is to say that we can't ever have real cognitive technologies because science will never understand intelligence. The master trick is unlike anything else known to science, therefore, it is magic. So at this point I would like to recite an inspirational quote from Lord Kelvin:

[Kelvin, from *Precise*.] "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... Modern biologists were coming once more to the acceptance of something and that was a vital principle."

Intelligence is not the first confusing phenomenon that science has ever encountered. The secret of life seemed a lot more mysterious, at the time, than intelligence seems to us now. Before the age of astronomy, stars seemed a lot more mysterious to astronomers than intelligence seems to us now. Likewise alchemy before the age of chemistry. People have no sense of historical perspective. They learn about stars and chemistry and biology in school and it seems that these matters have always been the proper meat of science, that they have *never been* mysterious. When science challenges some new puzzle, it is a great shock to the children of that generation, because they've never seen science successfully explain something that *feels* mysterious.

Over the last forty years there *has* been steady progress, believe it or not, in our basic understanding of the mind. There's even been steady progress in Artificial Intelligence. [AI creeping.] AI is slowly creeping up along the absolute scale of intelligence, but humans use the relative human scale, so all people see is that modern AIs are "dumber than a village idiot".

If you accept that the brain does not run on *elan vital*, and that the master trick is not magic, then at some point we're going to see some major cognitive technology. Something like a broadband brain-computer interface, or broadband computer-assisted telepathy that creates 64-node clustered humans, or genuine Artificial Intelligence. We're going to see cognitive technology that breaks the upper bound on intelligence which has held since the rise of the human species, something that cracks the ceiling of the known universe of mind.

And that is a sea change with the past, with a world of patterns created only by human intelligence. To understand the true strangeness of the future, we need to think about powerful intelligences that do not work the same way as modern human minds.

Anthropomorphism

[Human universals.] The anthropologist Donald Brown once compiled a list of more than 200 "human universals", characteristics that appear in every known human culture, from Palo Alto to Yanomamo hunter-gatherers in the Amazon rain forest. Tool making, weapons, grammar, tickling, meal times, mediating conflicts, dancing, singing, personal names, promises, mourning the dead. Anthropologists don't even think to report these characteristics because they're everywhere. 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! We forget how alike we are, under the skin, living in a world that reminds us only of our differences.

But remember: 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. The reason we humans share a common nature is a principle of evolutionary biology, which states

[Complex universal.] A complex adaptation, an adaptation with many interdependent genes, must be universal within a sexually reproducing species.

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.

[Genes.] 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. This won't happen often enough to give any of the genes a fitness advantage.

[Irreducible.] The way that complex machinery gets started is that a single gene, A, is an advantage all by itself, and it increases to universality. And then gene B which is dependent on that A arises and increases to universality. And then mutant A' which is dependent on B comes along, and increases to universality. And now you have "irreducible" complexity, A' that depends on B, and B that depends on A'. Then C comes along, which depends on A' and B, and then B' comes along, which depends on A' and C, and so on.

[X-Men] 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. That's not how evolution works. Natural selection has to follow an incremental path to complexity. At any given point in time, every member of the species has *some* version of the complex adaptation that's evolving.

When you apply this principle to the human mind, it gives rise to a rule that evolutionary psychologists have named [Unity] *the psychic unity of humankind*. Complex adaptations must be universal in a sexually reproducing species - including cognitive machinery in *Homo sapiens*. 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 thinking. 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 expression for disgust. This is the great failure of imagination - anthropomorphism.

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. 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 have to be wired the same way, therefore 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. If I am ignorant about a phenomenon, then this is a fact about my state of mind, not a fact about the phenomenon itself. Every mysterious phenomenon is mysterious to some particular person; there are no phenomena that are inherently mysterious. Including the brain, by the way.

[BEM.] Similarly, is a woman in a torn dress sexy, or is she sexy *to* human males? Since this woman is obviously *attractive*, of course the monster is *attracted* to her - isn't that logical? Your ancestors evolved in an environment where every mind they met worked the same way they did. So your brain has evolved to model other minds through empathy; asking "What would *I* do in this other mind's shoes?" This is going to give you *wrong* answers if the other mind does not work like you do. We can understand that not everyone believes the same things we do, because in the ancestral environment people did believe different things. But, as you can see from this painting here, we did not evolve to model swamp things with different emotions. So the empathy trick, using your own brain to model the other brain, gives the wrong answer. And the empathy trick is instinctive, unconscious, faster than deliberate thought. We don't think, "And now I'll use the empathy trick." We just think, "That girl is attractive, so the BEM will be attracted to her."

Minds in general

So if the empathy trick doesn't work, how can you predict other minds? Once we learn how to create true Artificial Intelligence, what will AIs be like? What will they do? This, of course, is the great, big, sixty-four trillion dollar, trick question.

In Hollywood movies, [Smiths] all the AIs are the same type, a single tribe. 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*. [Mindspace.] Imagine a map of mind design space. In one corner, a tiny little circle contains all humans. This is inside transhuman mindspace [Transhuman], which includes all

the human possibilities as a strict subset. This is inside posthuman mindspace [Posthuman], which is everything a transhuman might grow up into. And then there's all the rest of mind design space [AIs], the space of minds in general - including AIs so strange they aren't even recognizably posthuman. [You are here.]

Resist the temptation to generalize over all of mind design space. Let's say that the core of a mind in this space is embodied in a billion bits. Then there are two to the billionth power possible minds made up of a billion bits. And every universal generalization you try to make that covers all those minds has two to the billionth power chances to be false. You can't talk about "AIs" because you can't talk about all possible minds at once. First you have to point to somewhere specific in mindspace, say what kind of mind you're talking about. 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 an anthropomorphic, generally *wrong* answer. There's a lot more room in the category "AI" than there is in the category "human".

And this lesson applies beyond pure AIs. I suspect that even if you start with a human base and modify it - even if you're just working with brain hacks, or brain-computer interfaces, that sort of thing - you'll still get results a heck of a lot more surprising than a lot of people seem to expect. If you look at Earth right now, there's a lot of weird people out there, packed into the human dot. Step outside the human dot, and things will get stranger than they've ever been in *human* history. Unless, of course, that's not what the transhumans want.

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 create 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*. But then, most futurists are part of the entertainment industry - they tell wonderful stories about the future - and the first rule of storytelling is to be specific. A story, to be entertaining, cannot be vague; so if you don't know, you make something up. Saying "AIs will cure your cancer!" or "AIs will take over the world!" sounds much more interesting than saying "I don't know" or "It depends on the exact initial conditions of the AI."

Friendly AI

The Giant Cheesecake Fallacy is another case of trying to reason about the future by asking what "AIs" will "want", forgetting [Minds in general] that the space of possible minds is much wider than the tiny human dot.

In principle most people are willing to agree that other minds can be different. Until you present them with some particular concrete difference, and then they'll say it's impossible. Sooner or later they'll unconsciously use the empathy trick, use their own brain to simulate the other mind, and they'll believe the answer their own brain gives back is the only answer.

For example. I sometimes say that you should never trust a politician unless you have seen its source code. In our current world, there's no way to be sure that your elected representatives are working on your behalf, and often they aren't. So maybe we should replace the President with an open-source Artificial Intelligence [Prez AI], so we can all view the source code and make sure the AI really does care more about civil rights than holding on to power.

A typical response I get is that if you give power to an AI, the AI will become a tyrant - absolute power corrupts absolutely. But! Human beings have specifically evolved to be corrupted by power. When it comes to politics, human beings are defective by design. We evolved to deceive ourselves that we're taking over for the good of the tribe, and then use the political power to grab the best food and mates. There may even be a specific circuit in the brain somewhere that makes human beings want to abuse power, and it might be that all we need to get good government is to install circuit breakers on our politicians. It probably isn't that simple in real life, but, who knows, we haven't tried it, maybe it *is* that simple. Evolution adapted humans to respond to having power by abusing that power for personal interest. If I was building an AI from scratch, it would be just as easy to shape the AI so that, instead of responding to political power by being corrupted, it responded by using the power as little as possible.

[Mindspace.] Somewhere in the unimaginable vastness of mind design space, in fact, is a mind that responds to political power by singing a tune and doing a little dance. If you think *that* sounds odd, think of how strange it would look to an alien species that didn't have a sense of humor, if I smushed a cream pie into my face and the audience started laughing.

All previous political systems have suffered from the crippling handicap of being made entirely out of humans. If you have new raw materials, you can build new things. If you don't realize how wide mind design space is, you'll miss all sorts of interesting possibilities. Three hundred years ago, democracy was a wild-eyed crazy idea. The whole notion of having written laws is just a couple of thousand years old. The human species is just a few dozen millennia old, which is an

eyeblink of cosmological time. Things have not always been the way they are now. And they're not going to stay this way, either.

The scary truth is that there are no limits. With enough knowledge, you can build *any* kind of mind you can imagine. It is as free as the art of computer programming itself. There are as many possible minds as possible programs or possible configurations of atoms, because infinity equals infinity equals infinity. Admittedly, right now, we don't know enough to build any minds at all - neither augmented humans or pure AIs. This ignorance is a temporary condition. It won't be long before the absolute freedom starts. And then you're going to be faced with a choice. There are infinite possibilities, but which ones should we make real? What do you want to be [Choice] when you grow up? If there were no limits, who would you be? - because there aren't any! We are faced with absolute freedom - and rather than going into existentialist shock mode, I suggest that we deal with it and move on. I mean, all you need to do is figure out what you *want* out of life. That shouldn't be too hard, right?

RSI

There is a final point on the importance of cognitive technologies. Once upon a time, the way the world used to work [I->T] was that intelligence created technology. Brains made space shuttles and that was all. Now suppose that you've got someone with a brain-computer interface that makes them substantially more intelligent. They should be able to create more technology. What kind of technologies might they create? Flying cars? Cancer cures? One good bet is that they'd use their enhanced intelligence to create the next generation of brain-computer interfaces.

[I<->T] Cognitive technology closes the loop between intelligence and technology, creating a positive feedback cycle. The smarter you are, the more cognitive technology you can invent to make yourself even smarter. It's a tipping point - like a pen balanced on its point, once it tips over even a little, it soon falls the rest of the way.

[I. J. Good.] The purest form of this positive feedback cycle would be an AI rewriting its own source code, what the mathematician I. J. Good called an intelligence explosion. The AI redesigns itself and makes itself smarter, which increases the AI's ability to redesign itself, which allows the AI to rewrite itself yet again and make itself even smarter. Lather, rinse, repeat, FOOM.

I. J. Good's intelligence explosion should not be confused with Vernor Vinge's Singularity. Vernor Vinge's Singularity is the breakdown in our model of the future that occurs when we try to extrapolate that model past the point where the future contains entities smarter than us. I. J. Good's intelligence explosion should not be confused with Ray Kurzweil's accelerating change. The idea of an intelligence explosion does not logically require nor logically imply that 1970 to

2000 was a time of greater change than from 1940 to 1970. The first AI to improve itself could conceivably be created during an age of accelerating technological progress, or an age of slow but steady progress, or even a global stagnation of most technologies. And once we do go past the tipping point, it is not necessarily true that progress will be "exponential" - that's a specific mathematical curve, and it is not correct to say "exponential" when you mean "accelerating" or "positive second derivative".

Still, if you're wondering why the graph I showed earlier [AI arrow] only showed AI creeping up to the rough vicinity of human intelligence, and didn't show what happened after that, it's because an AI that is as smart as its programmers is smart enough to take over the job of improving itself. Not every positive feedback cycle in the universe follows a smooth exponential curve, but they do tend to accelerate.

For those of you in the audience who've never heard of me before and have no idea what the Singularity Institute for Artificial Intelligence does, then to make a very long story very short, we're trying to take humanity through the intelligence explosion safely by figuring out how to design a Friendly AI. [Friendly AI.] We want to reach into mind design space very precisely, with very exact targeting, and pull out an AI which is knowably nice, good, helpful, in a word, friendly, and which, furthermore, [Trajectory] stays that way while it improves itself. Somewhere in mind design space is an AI such that we will not regret creating it - but to find it will take new math; we'll have to actually know what we're doing instead of randomly messing around. New math takes a long time to develop, so we're getting started now. For more on this subject, see the Singularity Institute website at singinst.org.

[Sum up.] To sum up: 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. The rise of human intelligence has significantly transformed the world, we've carved our faces into mountains and set footprints on the moon. Whatever touches on intelligence reaches down to the roots and picks up the tree. A transhuman is a transhuman mind; anything else is a side issue. Cognitive technology opens up a vast space of possibilities unlike anything in human experience. It closes the loop and creates a positive feedback cycle. It may even let us create a self-improving AI, the fast road to superintelligence. If so, we need to work out new math in order to make that AI Friendly.

Above all, if you want to grasp where the future is going, ignore the cool devices with blinking lights, forget about mere superheros, and focus on technologies that impact upon the mind.

[Done.] This has been Eliezer Yudkowsky for the Singularity Institute for Artificial Intelligence.