

Transcript

The Human Importance of the Intelligence Explosion

The Singularity Summit at Stanford

May 13, 2006

Eliezer S. Yudkowsky

Talk

[Intro.] Good afternoon. I'm Eliezer Yudkowsky, a Research Fellow of the Singularity Institute for Artificial Intelligence. The title of today's talk is "The Human Importance of the Intelligence Explosion." [Explosion.] The term "intelligence explosion" was invented in 1965 by the eminent mathematician I. J. Good. The core idea of an intelligence explosion goes something like this: Suppose that we invent brain-computer interfaces that can substantially augment human intelligence. What will these augmented humans do with their improved intelligence? Develop new medical technologies? Play the stock market? One good bet is that they'll use their improved intelligence to design better brain-computer interfaces. And then, having become even smarter, they can invent even better brain-computer interfaces. Intelligence is the source of all technology, so if technology improves intelligence, that closes the loop and creates a positive feedback cycle. The purest case of this is a genuine AI, a fully intelligent AI, being able to rewrite its own source code, becoming smarter, and then rewriting its source code again. A recursively self-improving AI is what I. J. Good originally referred to as an "intelligence explosion" - although the notion generalizes.

[Distinct] I'd like to emphasize that the notion of an intelligence explosion is logically independent from many topics also linked to the word "Singularity". For example, 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. I emphasize this because many people mean different things by the word "Singularity", and if we talk about all these different things as if

they are the same, we will end up rather confused. That's one of the reasons I'm saying "intelligence explosion" rather than "Singularity" - "intelligence explosion" is more narrow and specific.

The title of this talk is "The human importance of the intelligence explosion." I intend to talk about intelligence first, then intelligence explosions, then finally the human importance of the intelligence explosion.

[Book smarts.] So, first, intelligence. When people hear the word "intelligence", they often think of what we might call "book smarts" - stereotypically "intelligent" activities like, calculus or chess. It takes more than book smarts to succeed in the human world: Enthusiasm, social skills, education, strategic cunning, rationality - but, note, each factor I listed is cognitive. Social manipulation is a talent that resides in the brain, not the kidneys.

[Parochial.] I suggest that we systematically underestimate the importance of intelligence, because in our everyday life, everyone we're dealing with is a human, as opposed to a mouse, a lizard, or a rock. We act like our tiny corner of mindspace is the whole world; we think of the scale of intelligence as if it ran from a village idiot to Einstein, rather than running from amoebas to humans. [Cosmopolitan.] On the scale of interspecies differences in intelligence, the entire human species fits in a small dot.

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 reshaped the Earth. The land sprouted skyscrapers, planes flew through the skies, footprints appeared on the Moon. If you look around you right now, most of the things you'll see are byproducts of human intelligence - the chairs, the floor, your clothes, all of these are effects with human intelligence as the cause.

[Power.] Intelligence is our trump card as a species, it's the human superpower. But we all have the same superpower so it doesn't seem very important. People say things like "intelligence is no match for a gun", as if guns had grown on trees. People say that intelligence doesn't count for as much as money, as if mice used money. All jokes aside, you won't find many great novelists, or military generals, or politicians, or scientists, who are lizards. Intelligence is the foundation of our power as a species, it's the strength that fuels our other arts.

[Unlike.] Advertisers want you to believe that the word "futuristic" means gleaming chrome and blinking lights, fascinating gadgets and expensive toys, because that's what they want to sell you. Imagine, if you like, that future biotechnology produces artificial red blood cells that let you hold your breath for four hours. I ask: So what? Humanity did not rise to prominence upon Earth by holding its breath longer than other species. The future technologies that will really matter are the technologies that impact upon the mind: Neurotechnology, brain-computer interfacing, and, of course, the purest form, Artificial Intelligence.

[Mystery] Intelligence is the most powerful force in the known universe, and also the most confusing question left in modern science. The phenomenon of intelligence is indisputably real - we see its effects every day - but you can ask "How does intelligence work?" and get ten different answers from ten different serious scientists. There's a great deal we know about the mind - it is *not* a complete mystery - but the knowledge is scattered across dozens of different fields - neuroanatomy, evolutionary psychology, computational neuroscience, information theory, cognitive psychology, Bayesian rationality, et cetera, et cetera, and of course the field currently called Artificial Intelligence. We know an enormous amount about the mind. But we also know there are things we don't yet know, because we have all this knowledge and we don't have human-level AI.

So there are things about the mind that we don't understand. [Jaynes] But let us remember the words of the mathematician E. T. Jaynes, who observed that if we are ignorant about a phenomenon, this is a fact about us, not a fact about the phenomenon itself. Confusion exists in our minds, not in reality. A blank spot on your map does not correspond to a blank territory. Science has encountered many mysterious questions, but we have yet to find a mysterious answer. So do not say, "intelligence is a mysterious phenomenon", because there are no phenomena which are mysterious in themselves. Rather, intelligence works in some completely ordinary way of which we are partially ignorant.

[More.] There's a *lot* more that I wish I could say about intelligence, but this is only a thirty-minute talk. The last time I gave a lecture on intelligence, I had ninety minutes to talk, and I still had to leave things out. If you google on "Singularity Institute" or go to singinst.org, and then click on the sidebar for "Summit Notes", you'll find a webpage that has a link to a video of the

ninety-minute lecture I gave on intelligence, and also some book chapters I've written for edited volumes.

Onward to I. J. Good's intelligence explosion. I originally gave the example of humans augmented with brain-computer interfaces, using their improved intelligence to build better brain-computer interfaces. [Bottleneck.] A difficulty with this scenario is that there's two parts to the system, the brain and the computer. If you want to improve the complete system, you can build interfaces with higher neural bandwidth to more powerful computers that do more cognitive work. But sooner or later you run into a bottleneck, which is the brain part of the brain-computer system. The core of your system has a serial speed of around a hundred operations per second. And worse, you can't reprogram it. Evolution did not build human brains to be hacked. Even if on the hardware level we could read and modify each individual neuron, and add neurons, and speed up neurons, we'd still be in trouble because the brain's software is a huge mess of undocumented spaghetti code. The human brain is not end-user-modifiable.

So trying to use brain-computer interfaces to create smarter-than-human intelligence may be like trying to build a heavier-than-air flying machine by strapping jet engines onto a bird. I'm not saying it could never, ever be done. *But* we might need a smarter-than-human AI just to handle the job of upgrading humans, *especially* if we want the upgrading process to be safe, sane, healthy, and pleasant. [Relative.] Upgrading humans may take a much more advanced technology, and a much more advanced understanding of cognitive science, than starting over and building a mind from scratch. Now, I do emphasize, that I am *totally* on board with the goal of humans becoming smarter. But I suspect that doing this gracefully will require molecular nanotechnology plus a very fast, very smart Artificial Intelligence that can customize strategies to your individual brain. Imagine if you had to take an existing, living bird, and expand the bird into a flyer the size of a 747, that actually flies, as fast as a 747, and do all this without killing the bird or making it very uncomfortable. Sufficiently advanced technology could do it, an ultrasmart AI with nanotech could do it; but you can't do it just by strapping jet engines onto the bird.

For similar reasons I don't think we should merge with our machines. My toaster is designed to toast bread, it is not designed to merge with me. Maybe I want my hands to generate enough heat to toast bread, so that I can cook breakfast through my own power, instead of relying on external

devices and outside crutches. Then it will take very advanced bioengineering to make my hands generate enough heat to toast bread, without damaging my hands or making them unable to do other things. There would be some optimal strategy for upgrading my hands in the healthiest, most graceful way, and that optimal strategy would not be to merge with a toaster.

20

[Advantage.] That's why recursive self-improvement and the intelligence explosion are usually discussed - I. J. Good originally discussed it - in the context of an AI rewriting its own source code. An AI can have total read and write access to its own state; move itself onto better hardware; get an explanation of its code from its programmers; work in the near-deterministic environment of a computer chip to make precise changes with knowable effects - in short, all the things we can't do with a human brain.

[Speed.] And an AI can be fast. In neuroscience there's a heuristic called the hundred-step rule, which is that anything a biological brain does in real time can require no more than one hundred serial computational steps. Whereas a modern computer chip can do a billion serial arithmetic operations per second. In the brain, the very fastest neurons transmit signals at 150 meters per second, which is two million times slower than the speed of light. Even on the issue of heat dissipation, where the brain does still beat modern transistors, the brain isn't anywhere near optimal. A synaptic spike in the brain dissipates more than a million times the minimum heat permitted by the laws of thermodynamics.

[Possible.] So it looks *physically* possible - that is, it's permitted by the known laws of physics - to have a computing device a million times faster than the brain, without shrinking the brain, running at ultra-low temperatures, dissipating more heat, or invoking quantum computing or reversible computing. You can even speed up the sensory inputs and motor outputs. Drexler's *Nanosystems* describes sensors and manipulators that operate on a timescale of nanoseconds - so again, a millionfold speedup appears to be permitted by the laws of physics. If a mind worked a million times as fast as a human brain, it would do a year of thinking in what, to us, would seem like 31 seconds. A century would go by in fifty-three minutes.

Let's say there's a smart AI that thinks a million times faster than a human, but the AI doesn't yet have tools that work on its own timescale. Starting from current human tools, what is the minimum time for the AI to bootstrap its capabilities to molecular nanotechnology? [Minimum.] I am only human, and I haven't spent all that much time thinking about the problem, so the fastest plausible method I can think of requires ten thousand years. First you need to crack the protein folding problem, not for biological proteins, but so that you can design chosen special cases of proteins. Then there are online providers that will accept an arbitrary DNA string, sequence the DNA, synthesize the peptides, and FedEx you the protein. Some providers boast of a 72-hour turnaround time. Email them the DNA sequence, you get the protein 72 hours later. So you order a set of proteins that, when mixed together, self-assemble into a very primitive nanomachine - something on the order of a ribosome, say - that can accept outside instructions in the form of coded acoustic vibrations. Then you use this very primitive nanodevice to build a second-stage nanomachine, that builds a third-stage nanomachine, and so on to full-scale molecular nanotechnology. This gives you nanotechnology in on the order of four days, which, at a millionfold speedup, is around ten thousand years of subjective time. Of course there might be faster ways to do it; I haven't spent ten thousand years thinking about the problem, and I'm only human.

Now this may or may not be roughly how things happen in real life. It probably won't be. The point I'm trying to make is to respect the power of creativity. [Power.] Intelligence isn't just about gadgets, even very powerful gadgets like nanomachines. Intelligence is about ingenuity and surprise, the strategy you didn't think of, going outside the box and changing the nature of the game. You might think, "Well, maybe the AI is millions of times smarter than a human, but what can it do if it can only manipulate the world through human tools?" You ought to think about that for at least a week, looking for a creative solution, before you put limits on a mind vastly smarter than us.

[Road.] Even saying that an ultrasmart mind might be able to bootstrap to nanotechnology in four days may still be a failure of imagination. Not just because there might be a more clever and creative way to do it even faster. But who says that molecular nanotechnology is the end of the line? The future has a reputation for accomplishing feats which the past thought impossible. If prophets of the year 1900 AD - never mind 1000 AD - had tried to bound the powers of human civilization a billion years later, some of

those physical impossibilities would have been accomplished before the century was out - transmuting lead into gold, for example. Because we remember future civilizations surprising past civilizations, people understand that we can't put limits on our great-great-grandchildren. And yet everyone in the 20th century, in the 19th century, and in the 11th century was human. The really scary part is when you start trying to imagine minds that are smarter than human in the same way that humans are smarter than mice. [Metaphors.] But even if you *just* talk about "An ultrafast mind with molecular nanotechnology", that's enough power to reshape the Earth, the Solar System, and the Milky Way.

If I say that an intelligence explosion might yield enough power to move every atom in the solar system to an exact new location, that may sound a little silly, a little ridiculous. If I say that it's enough power to take all the excess carbon dioxide out of the atmosphere and ship it to Mars, or alternatively, ship Earth's entire atmosphere to Pluto, it may sound a little far-fetched. But try convincing a hunter-gatherer from ten thousand years ago that someday people will invent an improved bow and arrow called a "nuclear missile" that wipes out everything in a ten-mile radius. Intelligence is ridiculously powerful. Literally so, in the sense that, over and over again, intelligence has accomplished things that once sounded ridiculous. And we're not talking about a jump like the one from hunter-gatherer society to the Internet; that's merely ten thousand years of subjective time with no differences of brain architecture. The intelligence explosion would be a much bigger jump. It could and probably will reshape the entire world.

[Avoidable.] Can we prevent an intelligence explosion from ever happening? This seems pretty unlikely to me. If you try to balance a pen exactly on its tip, it's very difficult because if the pen tilts a little, gravity pulls it over a little more, and the process accelerates. In the same sense, if technology improves intelligence a little, it becomes easier to invent even more powerful cognitive technologies, and again the process accelerates. Alternatively, a civilization can wipe itself out - that would be a stable state, the ashes of a dead planet forever circling the Sun. It seems to me that sooner or later, a civilization is bound to wander into one region or the other - a superintelligent region, or a dead planet region.

So what can we do? Because, of course, you're going to do *something*. You're not going to hear that the destiny of all humankind hangs in the balance, say, "Wow, cool, this was such an entertaining day," and then go

back to watching television. Other people might be that silly, but you're smarter than that. Right? But what *can* we do?

Well, there's certainly more than one possible way that humanity could go extinct. Maybe there's more than one region of superintelligence to wander into?

[Cheesecake.] One often hears, in futurism, a line of reasoning that goes something like this. 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 the argument leaps directly from *capability* to *actuality*, without considering the necessary intermediate of *motive*.

[Missing.] Here are two examples 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.]
- Or: 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.

And the natural *mistake*, once you understand the Giant Cheesecake Fallacy, is to ask: "What will an Artificial Intelligence want?"

[Mind design space.] When trying to talk about Artificial Intelligence, it becomes extremely important to remember that we cannot make any general statement about Artificial Intelligences because the design space is too large. People talk about "AIs" as if all AIs formed a single tribe, an ethnic stereotype. Now, it might make sense to talk about "the human species" as a natural category, because we humans all have essentially the same brain architecture - limbic system, cerebellum, visual cortex, prefrontal cortex, and so on. But the term "Artificial Intelligence" refers to a vastly larger *space of*

possibilities than this. When we talk about "AIs" we are really talking about minds-in-general. Imagine a map of mind design space. In one corner, a tiny little circle contains all humans. And then all the rest of that huge sphere is the space of minds in general.

10

Everyday experience tells us that humans often say and do many different things. And yet some futurists predict what "AIs" will do far more confidently than they would ever dare predict a specific, real human being. Is the space of minds smaller than the space of humans?

The error I keep seeing is that people try to generalize over all possible minds. Imagine that this sphere is a quintillion bits wide and contains two to the quintillionth power possible minds, including every possible human. Then an assertion that's supposedly true of *every mind in the sphere* has two to the quintillionth power chances to be falsified, two to the quintillionth power chances to be wrong. But if you say that at least one mind in the sphere has some property, then you have two to the quintillionth power chances to be right. Somewhere in this sphere is a mind that does not wish to kill human beings. I know, because I am in this sphere, and I don't want to kill people. If I could modify my own source code, I wouldn't knowingly modify myself to want to kill people, because people might die as a result of that, and my current self doesn't want people to die. Gandhi does not want to commit murder and does not want to modify himself to commit murder. That's a rather informal argument that a self-improving mind can have stable motivations. And, again, for more detailed discussion, check the Summit Notes on the Singularity Institute website.

[Engineering.] The larger point is that it's not a question of *predicting* what AIs *will do*. It's not a prediction problem; it's not a question of futurism; it's a challenge of engineering. If you have the knowledge, if you know exactly what you're doing, then you ought to be able to *reach* into mind design space, and pull out a mind such that we're glad we made it real. [Technical.] To say that this has to be done very, very carefully, vastly understates the seriousness of the matter.

If I may, I would like to deliver a dire warning that is also useful advice in general. [Warning.] Norman R. F. Maier discovered that if you instruct a group "Do not propose solutions until the problem has been discussed as

thoroughly as possible without suggesting any", then that group produces better solutions than groups not given this instruction. Robyn Dawes added that when groups face a very tough problem, that is when they are most likely to propose solutions immediately. Making the right choice in creating AI is such an *incredibly difficult* problem, with so many different facets and challenges and requirements and ways to fail, that it is impossible to discuss it with people because they *instantly* come up with a solution. Or they instantly declare the problem unsolvable. Either way it's the same mistake.

[What kind?] Trying to describe what kind of AI we want to see is a very tricky business. It's a lot easier to describe AIs we don't want to see. Science fiction and Hollywood movies have lots of AIs we don't want to see. Fiction is about conflict. Conflict requires problems. People sometimes ask me whether Isaac Asimov's Three Laws of Robotics are a good solution. Leaving aside the extensive technical difficulties, if Asimov had depicted robots that actually worked correctly, he would have had no story. "Everything worked fine, the end." Asimov *chose* his Three Laws to create story conflicts. Similarly, it's a common theme in fiction that we'll build AI, and the AI will be so much better at everything humans do that the AI will take all the jobs, and humans will starve, or sit around watching television all day. And yes, I would view that as a sad ending, a waste of what the human species could have been. Doing someone's work for them is not always helpful. But it is a Giant Cheesecake Fallacy to say that just because an AI *can* do a job, the AI *will* do that job. Some possible minds might decide to do all our work for us, but some minds won't. So if we end up watching television all day, or feeling that our lives and destinies have been taken away from us and we've become nothing but pets, it's because someone pulled the wrong mind out of design space. Similarly, of course, if we all end up dead.

I think that sad endings are possible, but not inevitable. I think we can find a happy beginning, *if we know exactly what we're doing*. I don't think we can win if we're poking around randomly in mind design space, with no idea what's going to pop out. Very few people want to destroy the world, very few people would create hell on earth on purpose, but good intentions can't make up for ignorance or carelessness. [Friendly.] I use the term "Friendly AI" to refer to this whole challenge, creating a mind that doesn't kill people and does cure cancer, the problem of pulling a mind out of design space such that afterwards you're glad you did it. The challenge of Friendly AI, looks

possible but *very difficult*, and the whole subject is *extremely* easy to misunderstand if you start jumping to conclusions.

Nonetheless, I think we can win *if we try*. And, if we do win, then the world *can* become a better place. [Better.] An intelligence explosion is a lot of power, and that power doesn't *have* to make things worse. There are things I want to do for myself, rather than an AI doing them for me - but there are also problems in the world that are so awful that it is more important to solve them as quickly as possible than that we solve them for ourselves. I don't think there's anything philosophically unacceptable about a superintelligence curing cancer or AIDS. To put it in the nicest possible way, this planet has lots and lots and lots of room for improvement. Which *is* a nice way of putting it. It doesn't say that you have to be gloomy today, but even the things that are already good, can get better. I think that right now the human species is going through this awkward adolescent stage when we're smart enough to make huge problems for ourselves but not quite smart enough to solve them. [Someday.] Someday the human species has to grow up. Why not sooner rather than later? That's the scary and beautiful thing about increased intelligence. It's *not* just another shiny, expensive gadget. Increased intelligence might let us make real progress on our deepest problems.

The Milky Way has four hundred billion stars, and in the visible universe there are sixty billion galaxies. So far as we know, we are alone. Maybe intelligent life is just very improbable. Someone has to be first. But you can look up at the night sky and see how much work there is for humankind to do. Look around you at this world, in all its beauty and all its ugliness. Is this where we stop and declare that our work is finished? I don't think so. Not with so many people in pain, and so many people living lives of quiet desperation, and not with all those stars twinkling in the night sky.

In this world, there are well-paid, highly trained professionals whose job it is to sell lipstick, because the owners of the lipstick corporation care too much about selling lipstick to leave the job to amateurs or unpaid volunteers. The Singularity Institute was founded on the principle that the fate of the human species deserves at least one-thousandth the consideration we give to lipstick. Someday after all of this is over, an awful lot of people are going to look back and kick themselves and say, "What on Earth was I doing? Why was I sitting around watching television while the fate of all humankind hung in the balance and almost no one was doing anything?" [Remember.]

In a hundred million years, no one's going to care who won the World Series, but they'll remember the first AI.

So let us raise the banner and sound the trumpets, fund the grants and do the research, and see if the human species can give this problem one-thousandth of the resources that we expend on lipstick. Thank you.

[More.] For a lot more information, and all the parts I had to leave out of this talk, please visit the Singularity Institute website at singinst.org, or google on Singularity Institute. The Singularity Institute for Artificial Intelligence is a 501(c)(3) nonprofit organization, funded by individual donations from people *just like you*. If you all go back to your day jobs with a warm fuzzy feeling and a sense of renewed meaning, but you don't actually do anything as a result, then this day has accomplished nothing. Intelligence is our species superpower, but intelligence, to be useful, must actually be used. If you're already spending *every dollar*, on matters of drastic importance to the fate of a billion galaxies, forget I said anything. Otherwise, we do need your help, not just your applause. Thank you.