ABBILDUNGEN AUS DER

# METEOROLOGIE,

DIE VERSCHIEDENEN ERSCHEINUNGEN DER ATMOSPHÄRE DARSTELLEND.

# Futurism and Forecasting

1. Superintelligence FAQ
2. AI Researchers On AI Risk
3. Should AI Be Open?
4. SSC Journal Club: AI Timelines
5. Where The Falling Einstein Meets The Rising Mouse
6. Don’t Fear The Filter
7. Book Review: Age of Em
8. Ascended Economy?
9. G.K. Chesterton On AI Risk
10. [REPOST] The Demiurge’s Older Brother

# Superintelligence FAQ

### 1: What is superintelligence?

A superintelligence is a mind that is much more intelligent than any human. Most of the time, it's used to discuss hypothetical future AIs.

### 1.1: Sounds a lot like science fiction. Do people think about this in the real world?

Yes. Two years ago, Google bought artificial intelligence startup DeepMind for $400 million; DeepMind added the condition that Google promise to set up an AI Ethics Board. DeepMind cofounder Shane Legg has said in interviews that he believes superintelligent AI will be "something approaching absolute power" and "the number one risk for this century".

Many other science and technology leaders agree. Astrophysicist Stephen Hawking says that superintelligence "could spell the end of the human race." Tech billionaire Bill Gates describes himself as "in the camp that is concerned about superintelligence…I don't understand why some people are not concerned". SpaceX/Tesla CEO Elon Musk calls superintelligence "our greatest existential threat" and donated $10 million from his personal fortune to study the danger. Stuart Russell, Professor of Computer Science at Berkeley and world-famous AI expert, warns of "species-ending problems" and wants his field to pivot to make superintelligence-related risks a central concern.

Professor Nick Bostrom is the director of Oxford's Future of Humanity Institute, tasked with anticipating and preventing threats to human civilization. He has been studying the risks of artificial intelligence for twenty years. The explanations below are loosely adapted from his 2014 book Superintelligence, and divided into three parts addressing three major questions. First, why is superintelligence a topic of concern? Second, what is a "hard takeoff" and how does it impact our concern about superintelligence? Third, what measures can we take to make superintelligence safe and beneficial for humanity?

### 2: AIs aren't as smart as rats, let alone humans. Isn't it sort of early to be worrying about this kind of thing?

Maybe. It's true that although AI has had some recent successes – like DeepMind's newest creation AlphaGo defeating the human Go champion in April – it still has nothing like humans' flexible, cross-domain intelligence. No AI in the world can pass a first-grade reading comprehension test. Facebook's Andrew Ng compares worrying about superintelligence to "worrying about overpopulation on Mars" – a problem for the far future, if at all.

But this apparent safety might be illusory. A survey of leading AI scientists show that on average they expect human-level AI as early as 2040, with above-human-level AI following shortly after. And many researchers warn of a possible "fast takeoff" – a point around human-level AI where progress reaches a critical mass and then accelerates rapidly and unpredictably.

### 2.1: What do you mean by "fast takeoff"?

A slow takeoff is a situation in which AI goes from infrahuman to human to superhuman intelligence very gradually. For example, imagine an augmented "IQ" scale (THIS IS NOT HOW IQ ACTUALLY WORKS – JUST AN EXAMPLE) where rats weigh in at 10, chimps at 30, the village idiot at 60, average humans at 100, and Einstein at 200. And suppose that as technology advances, computers gain two points on this scale per year. So if they start out as smart as rats in 2020, they'll be as smart as chimps in 2035, as smart as the village idiot in 2050, as smart as average humans in 2070, and as smart as Einstein in 2120. By 2190, they'll be IQ 340, as far beyond Einstein as Einstein is beyond a village idiot.

In this scenario progress is gradual and manageable. By 2050, we will have long since noticed the trend and predicted we have 20 years until average-human-level intelligence. Once AIs reach average-human-level intelligence, we will have fifty years during which some of us are still smarter than they are, years in which we can work with them as equals, test and retest their programming, and build institutions that promote cooperation. Even though the AIs of 2190 may qualify as "superintelligent", it will have been long-expected and there would be little point in planning now when the people of 2070 will have so many more resources to plan with.

A moderate takeoff is a situation in which AI goes from infrahuman to human to superhuman relatively quickly. For example, imagine that in 2020 AIs are much like those of today – good at a few simple games, but without clear domain-general intelligence or "common sense". From 2020 to 2050, AIs demonstrate some academically interesting gains on specific problems, and become better at tasks like machine translation and self-driving cars, and by 2047 there are some that seem to display some vaguely human-like abilities at the level of a young child. By late 2065, they are still less intelligent than a smart human adult. By 2066, they are far smarter than Einstein.

A fast takeoff scenario is one in which computers go even faster than this, perhaps moving from infrahuman to human to superhuman in only days or weeks.

### 2.1.1: Why might we expect a moderate takeoff?

Because this is the history of computer Go, with fifty years added on to each date. In 1997, the best computer Go program in the world, Handtalk, won NT$250,000 for performing a previously impossible feat – beating an 11 year old child (with an 11-stone handicap penalizing the child and favoring the computer!) As late as September 2015, no computer had ever beaten any professional Go player in a fair game. Then in March 2016, a Go program beat 18-time world champion Lee Sedol 4-1 in a five game match. Go programs had gone from "dumber than children" to "smarter than any human in the world" in eighteen years, and "from never won a professional game" to "overwhelming world champion" in six months.

The slow takeoff scenario mentioned above is loading the dice. It theorizes a timeline where computers took fifteen years to go from "rat" to "chimp", but also took thirty-five years to go from "chimp" to "average human" and fifty years to go from "average human" to "Einstein". But from an evolutionary perspective this is ridiculous. It took about fifty million years (and major redesigns in several brain structures!) to go from the first rat-like creatures to chimps. But it only took about five million years (and very minor changes in brain structure) to go from chimps to humans. And going from the average human to Einstein didn't even require evolutionary work – it's just the result of random variation in the existing structures!

So maybe our hypothetical IQ scale above is off. If we took an evolutionary and neuroscientific perspective, it would look more like flatworms at 10, rats at 30, chimps at 60, the village idiot at 90, the average human at 98, and Einstein at 100.

Suppose that we start out, again, with computers as smart as rats in 2020. Now we get still get computers as smart as chimps in 2035. And we still get computers as smart as the village idiot in 2050. But now we get computers as smart as the average human in 2054, and computers as smart as Einstein in 2055. By 2060, we're getting the superintelligences as far beyond Einstein as Einstein is beyond a village idio

This offers a much shorter time window to react to AI developments. In the slow takeoff scenario, we figured we could wait until computers were as smart as humans before we had to start thinking about this; after all, that still gave us fifty years before computers were even as smart as Einstein. But in the moderate takeoff scenario, it gives us one year until Einstein and six years until superintelligence. That's starting to look like not enough time to be entirely sure we know what we're doing.

### 2.1.2: Why might we expect a fast takeoff?

AlphaGo used about 0.5 petaflops (= trillion floating point operations per second) in its championship game. But the world's fastest supercomputer, TaihuLight, can calculate at almost 100 petaflops. So suppose Google developed a human-level AI on a computer system similar to AlphaGo, caught the attention of the Chinese government (who run TaihuLight), and they transfer the program to their much more powerful computer. What would happen?

It depends on to what degree intelligence benefits from more computational resources. This differs for different processes. For domain-general intelligence, it seems to benefit quite a bit – both across species and across human individuals, bigger brain size correlates with greater intelligence. This matches the evolutionarily rapid growth in intelligence from chimps to hominids to modern man; the few hundred thousand years since australopithecines weren't enough time to develop complicated new algorithms, and evolution seems to have just given humans bigger brains and packed more neurons and glia in per square inch. It's not really clear why the process stopped (if it ever did), but it might have to do with heads getting too big to fit through the birth canal. Cancer risk might also have been involved – scientists have found that smarter people are more likely to get brain cancer, possibly because they're already overclocking their ability to grow brain cells.

At least in neuroscience, once evolution "discovered" certain key insights, further increasing intelligence seems to have been a matter of providing it with more computing power. So again – what happens when we transfer the hypothetical human-level AI from AlphaGo to a TaihuLight-style supercomputer two hundred times more powerful? It might be a stretch to expect it to go from IQ 100 to IQ 20,000, but might it increase to an Einstein-level 200, or a superintelligent 300? Hard to say – but if Google ever does develop a human-level AI, the Chinese government will probably be interested in finding out.

Even if its intelligence doesn't scale linearly, TaihuLight could give it more time. TaihuLight is two hundred times faster than AlphaGo. Transfer an AI from one to the other, and even if its intelligence didn't change – even if it had exactly the same thoughts – it would think them two hundred times faster. An Einstein-level AI on AlphaGo hardware might (like the historical Einstein) discover one revolutionary

breakthrough every five years. Transfer it to TaihuLight, and it would work two hundred times faster – a revolutionary breakthrough every week.

Supercomputers track Moore's Law; the top supercomputer of 2016 is a hundred times faster than the top supercomputer of 2006. If this progress continues, the top computer of 2026 will be a hundred times faster still. Run Einstein on that computer, and he will come up with a revolutionary breakthrough every few hours. Or something. At this point it becomes a little bit hard to imagine. All I know is that it only took one Einstein, at normal speed, to lay the theoretical foundation for nuclear weapons. Anything a thousand times faster than that is definitely cause for concern.

There's one final, very concerning reason to expect a fast takeoff. Suppose, once again, we have an AI as smart as Einstein. It might, like the historical Einstein, contemplate physics. Or it might contemplate an area very relevant to its own interests: artificial intelligence. In that case, instead of making a revolutionary physics breakthrough every few hours, it will make a revolutionary AI breakthrough every few hours. Each AI breakthrough it makes, it will have the opportunity to reprogram itself to take advantage of its discovery, becoming more intelligent, thus speeding up its breakthroughs further. The cycle will stop only when it reaches some physical limit – some technical challenge to further improvements that even an entity far smarter than Einstein cannot discover a way around.

To human programmers, such a cycle would look like a "critical mass". Before the critical level, any AI advance delivers only modest benefits. But any tiny improvement that pushes an AI above the critical level would result in a feedback loop of inexorable self-improvement all the way up to some stratospheric limit of possible computing power.

This feedback loop would be exponential; relatively slow in the beginning, but blindingly fast as it approaches an asymptote. Consider the AI which starts off making forty breakthroughs per year – one every nine days. Now suppose it gains on average a 10% speed improvement with each breakthrough. It starts on January 1. Its first breakthrough comes January 10 or so. Its second comes a little faster, January 18. Its third is a little faster still, January 25. By the beginning of February, it's sped up to producing one breakthrough every seven days, more or less. By the beginning of March, it's making about one breakthrough every three days or so. But by March 20, it's up to one breakthrough a day. By late on the night of March 29, it's making a breakthrough every second.

#### 2.1.2.1: Is this just following an exponential trend line off a cliff?

This is certainly a risk (affectionately known in AI circles as "pulling a Kurzweill"), but sometimes taking an exponential trend seriously is the right response.

Consider economic doubling times. In 1 AD, the world GDP was about $20 billion; it took a thousand years, until 1000 AD, for that to double to $40 billion. But it only took five hundred more years, until 1500, or so, for the economy to double again. And then it only took another three hundred years or so, until 1800, for the economy to double a third time. Someone in 1800 might calculate the trend line and say this was ridiculous, that it implied the economy would be doubling every ten years or so in the beginning of the 21st century. But in fact, this is how long the economy takes to double these days. To a medieval, used to a thousand-year doubling time (which was based mostly on population growth!), an economy that doubled every ten years might seem inconceivable. To us, it seems normal.

Likewise, in 1965 Gordon Moore noted that semiconductor complexity seemed to double every eighteen months. During his own day, there were about five hundred transistors on a chip; he predicted that would soon double to a thousand, and a few years later to two thousand. Almost as soon as Moore's Law become well-known, people started saying it was absurd to follow it off a cliff – such a law would imply a million transistors per chip in 1990, a hundred million in 2000, ten billion transistors on every chip by 2015! More transistors on a single chip than existed on all the computers in the world! Transistors the size of molecules! But of course all of these things happened; the ridiculous exponential trend proved more accurate than the naysayers.

None of this is to say that exponential trends are always right, just that they are sometimes right even when it seems they can't possibly be. We can't be sure that a computer using its own intelligence to discover new ways to increase its intelligence will enter a positive feedback loop and achieve superintelligence in seemingly impossibly short time scales. It's just one more possibility, a worry to place alongside all the other worrying reasons to expect a moderate or hard takeoff.

### 2.2: Why does takeoff speed matter?

A slow takeoff over decades or centuries would give us enough time to worry about superintelligence during some indefinite "later", making current planning as silly as worrying about "overpopulation on Mars". But a moderate or hard takeoff means there wouldn't be enough time to deal with the problem as it occurs, suggesting a role for preemptive planning.

(in fact, let's take the "overpopulation on Mars" comparison seriously. Suppose Mars has a carrying capacity of 10 billion people, and we decide it makes sense to worry about overpopulation on Mars only once it is 75% of the way to its limit. Start with 100 colonists who double every twenty years. By the second generation there are 200 colonists; by the third, 400. Mars reaches 75% of its carrying capacity after 458 years, and crashes into its population limit after 464 years. So there were 464 years in which the Martians could have solved the problem, but they insisted on waiting until there were only six years left. Good luck solving a planetwide population crisis in six years. The moral of the story is that exponential trends move faster than you think and you need to start worrying about them early).

## 3: Why might a fast takeoff be dangerous?

The argument goes: yes, a superintelligent AI might be far smarter than Einstein, but it's still just one program, sitting in a supercomputer somewhere. That could be bad if an enemy government controls it and asks its help inventing superweapons – but then the problem is the enemy government, not the AI per se. Is there any reason to be afraid of the AI itself? Suppose the AI did feel hostile – suppose it even wanted to take over the world? Why should we think it has any chance of doing so?

Compounded over enough time and space, intelligence is an awesome advantage. Intelligence is the only advantage we have over lions, who are otherwise much bigger and stronger and faster than we are. But we have total control over lions, keeping them in zoos to gawk at, hunting them for sport, and holding them on the brink of extinction. And this isn't just the same kind of quantitative advantage tigers have over lions, where maybe they're a little bigger and stronger but they're at least on a level playing field and enough lions could probably overpower the tigers. Humans are playing a completely different game than the lions, one that no lion will ever be able

to respond to or even comprehend. Short of human civilization collapsing or lions evolving human-level intelligence, our domination over them is about as complete as it is possible for domination to be.

Since superintelligences will be as far beyond Einstein as Einstein is beyond a village idiot, we might worry that they would have the same kind of qualitative advantage over us that we have over lions.

**3.1: Human civilization as a whole is dangerous to lions. But a single human placed amid a pack of lions with no raw materials for building technology is going to get ripped to shreds. So although thousands of superintelligences, given a long time and a lot of opportunity to build things, might be able to dominate humans – what harm could a single superintelligence do?**

Superintelligence has an advantage that a human fighting a pack of lions doesn't – the entire context of human civilization and technology, there for it to manipulate socially or technologically.

**3.1.1: What do you mean by superintelligences manipulating humans socially?**

People tend to imagine AIs as being like nerdy humans – brilliant at technology but clueless about social skills. There is no reason to expect this – persuasion and manipulation is a different kind of skill from solving mathematical proofs, but it's still a skill, and an intellect as far beyond us as we are beyond lions might be smart enough to replicate or exceed the "charming sociopaths" who can naturally win friends and followers despite a lack of normal human emotions. A superintelligence might be able to analyze human psychology deeply enough to understand the hopes and fears of everyone it negotiates with. Single humans using psychopathic social manipulation have done plenty of harm – Hitler leveraged his skill at oratory and his understanding of people's darkest prejudices to take over a continent. Why should we expect superintelligences to do worse than humans far less skilled than they?

(More outlandishly, a superintelligence might just skip language entirely and figure out a weird pattern of buzzes and hums that causes conscious thought to seize up, and which knocks anyone who hears it into a weird hypnotizable state in which they'll do anything the superintelligence asks. It sounds kind of silly to me, but then, nuclear weapons probably would have sounded kind of silly to lions sitting around speculating about what humans might be able to accomplish. When you're dealing with something unbelievably more intelligent than you are, you should probably expect the unexpected.)

**3.1.2: What do you mean by superintelligences manipulating humans technologically?**

AlphaGo was connected to the Internet – why shouldn't the first superintelligence be? This gives a sufficiently clever superintelligence the opportunity to manipulate world computer networks. For example, it might program a virus that will infect every computer in the world, causing them to fill their empty memory with partial copies of the superintelligence, which when networked together become full copies of the superintelligence. Now the superintelligence controls every computer in the world, including the ones that target nuclear weapons. At this point it can force humans to bargain with it, and part of that bargain might be enough resources to establish its own industrial base, and then we're in humans vs. lions territory again.

(Satoshi Nakamoto is a mysterious individual who posted a design for the Bitcoin currency system to a cryptography forum. The design was so brilliant that everyone started using it, and Nakamoto – who had made sure to accumulate his own store of the currency before releasing it to the public – became a multibillionaire. In other words, somebody with no resources except the ability to make one post to an Internet forum managed to leverage that into a multibillion dollar fortune – and he wasn't even superintelligent. If Hitler is a lower-bound on how bad superintelligent persuaders can be, Nakamoto should be a lower-bound on how bad superintelligent programmers with Internet access can be.)

### 3.2: Couldn't sufficiently paranoid researchers avoid giving superintelligences even this much power?

That is, if you know an AI is likely to be superintelligent, can't you just disconnect it from the Internet, not give it access to any speakers that can make mysterious buzzes and hums, make sure the only people who interact with it are trained in caution, et cetera?. Isn't there some level of security – maybe the level we use for that room in the CDC where people in containment suits hundreds of feet underground analyze the latest superviruses – with which a superintelligence could be safe?

This puts us back in the same situation as lions trying to figure out whether or not nuclear weapons are a things humans can do. But suppose there is such a level of security. You build a superintelligence, and you put it in an airtight chamber deep in a cave with no Internet connection and only carefully-trained security experts to talk to. What now?

Now you have a superintelligence which is possibly safe but definitely useless. The whole point of building superintelligences is that they're smart enough to do useful things like cure cancer. But if you have the monks ask the superintelligence for a cancer cure, and it gives them one, that's a clear security vulnerability. You have a superintelligence locked up in a cave with no way to influence the outside world except that you're going to mass produce a chemical it gives you and inject it into millions of people.

Or maybe none of this happens, and the superintelligence sits inert in its cave. And then another team somewhere else invents a second superintelligence. And then a third team invents a third superintelligence. Remember, it was only about ten years between Deep Blue beating Kasparov, and everybody having Deep Blue – level chess engines on their laptops. And the first twenty teams are responsible and keep their superintelligences locked in caves with carefully-trained experts, and the twenty-first team is a little less responsible, and now we still have to deal with a rogue superintelligence.

Superintelligences are extremely dangerous, and no normal means of controlling them can entirely remove the danger.

### 4: Even if hostile superintelligences are dangerous, why would we expect a superintelligence to ever be hostile?

The argument goes: computers only do what we command them; no more, no less. So it might be bad if terrorists or enemy countries develop superintelligence first. But if we develop superintelligence first there's no problem. Just command it to do the things we want, right?

Suppose we wanted a superintelligence to cure cancer. How might we specify the goal "cure cancer"? We couldn't guide it through every individual step; if we knew every individual step, then we could cure cancer ourselves. Instead, we would have to give it a final goal of curing cancer, and trust the superintelligence to come up with intermediate actions that furthered that goal. For example, a superintelligence might decide that the first step to curing cancer was learning more about protein folding, and set up some experiments to investigate protein folding patterns.

A superintelligence would also need some level of common sense to decide which of various strategies to pursue. Suppose that investigating protein folding was very likely to cure 50% of cancers, but investigating genetic engineering was moderately likely to cure 90% of cancers. Which should the AI pursue? Presumably it would need some way to balance considerations like curing as much cancer as possible, as quickly as possible, with as high a probability of success as possible.

But a goal specified in this way would be very dangerous. Humans instinctively balance thousands of different considerations in everything they do; so far this hypothetical AI is only balancing three (least cancer, quickest results, highest probability). To a human, it would seem maniacally, even psychopathically, obsessed with cancer curing. If this were truly its goal structure, it would go wrong in almost comical ways.

If your only goal is "curing cancer", and you lack humans' instinct for the thousands of other important considerations, a relatively easy solution might be to hack into a nuclear base, launch all of its missiles, and kill everyone in the world. This satisfies all the AI's goals. It reduces cancer down to zero (which is better than medicines which work only some of the time). It's very fast (which is better than medicines which might take a long time to invent and distribute). And it has a high probability of success (medicines might or might not work; nukes definitely do).

So simple goal architectures are likely to go very wrong unless tempered by common sense and a broader understanding of what we do and do not value.

**4.1: But superintelligences are very smart. Aren't they smart enough not to make silly mistakes in comprehension?**

Yes, a superintelligence should be able to figure out that humans will not like curing cancer by destroying the world. However, in the example above, the superintelligence is programmed to follow human commands, not to do what it thinks humans will "like". It was given a very specific command – cure cancer as effectively as possible. The command makes no reference to "doing this in a way humans will like", so it doesn't.

(by analogy: we humans are smart enough to understand our own "programming". For example, we know that – pardon the anthromorphizing – evolution gave us the urge to have sex so that we could reproduce. But we still use contraception anyway. Evolution gave us the urge to have sex, not the urge to satisfy evolution's values directly. We appreciate intellectually that our having sex while using condoms doesn't carry out evolution's original plan, but – not having any particular connection to evolution's values – we don't care)

We started out by saying that computers only do what you tell them. But any programmer knows that this is precisely the problem: computers do exactly what you tell them, with no common sense or attempts to interpret what the instructions really meant. If you tell a human to cure cancer, they will instinctively understand how this

interacts with other desires and laws and moral rules; if you tell an AI to cure cancer, it will literally just want to cure cancer.

Define a closed-ended goal as one with a clear endpoint, and an open-ended goal as one to do something as much as possible. For example “find the first one hundred digits of pi” is a closed-ended goal; “find as many digits of pi as you can within one year” is an open-ended goal. According to many computer scientists, giving a superintelligence an open-ended goal without activating human instincts and counterbalancing considerations will usually lead to disaster.

To take a deliberately extreme example: suppose someone programs a superintelligence to calculate as many digits of pi as it can within one year. And suppose that, with its current computing power, it can calculate one trillion digits during that time. It can either accept one trillion digits, or spend a month trying to figure out how to get control of the TaihuLight supercomputer, which can calculate two hundred times faster. Even if it loses a little bit of time in the effort, and even if there’s a small chance of failure, the payoff – two hundred trillion digits of pi, compared to a mere one trillion – is enough to make the attempt. But on the same basis, it would be even better if the superintelligence could control every computer in the world and set it to the task. And it would be better still if the superintelligence controlled human civilization, so that it could direct humans to build more computers and speed up the process further.

Now we’re back at the situation that started Part III – a superintelligence that wants to take over the world. Taking over the world allows it to calculate more digits of pi than any other option, so without an architecture based around understanding human instincts and counterbalancing considerations, even a goal like “calculate as many digits of pi as you can” would be potentially dangerous.

**5: Aren’t there some pretty easy ways to eliminate these potential problems?**

There are many ways that look like they can eliminate these problems, but most of them turn out to have hidden difficulties.

**5.1: Once we notice that the superintelligence working on calculating digits of pi is starting to try to take over the world, can’t we turn it off, reprogram it, or otherwise correct its mistake?**

No. The superintelligence is now focused on calculating as many digits of pi as possible. Its current plan will allow it to calculate two hundred trillion such digits. But if it were turned off, or reprogrammed to do something else, that would result in it calculating zero digits. An entity fixated on calculating as many digits of pi as possible will work hard to prevent scenarios where it calculates zero digits of pi. Indeed, it will interpret such as a hostile action. Just by programming it to calculate digits of pi, we will have given it a drive to prevent people from turning it off.

University of Illinois computer scientist Steve Omohundro argues that entities with very different final goals – calculating digits of pi, curing cancer, helping promote human flourishing – will all share a few basic ground-level subgoals. First, self-preservation – no matter what your goal is, it’s less likely to be accomplished if you’re too dead to work towards it. Second, goal stability – no matter what your goal is, you’re more likely to accomplish it if you continue to hold it as your goal, instead of going off and doing something else. Third, power – no matter what your goal is, you’re more likely to be able to accomplish it if you have lots of power, rather than very little.

So just by giving a superintelligence a simple goal like “calculate digits of pi”, we’ve accidentally given it Omohundro goals like “protect yourself”, “don’t let other people reprogram you”, and “seek power”.

As long as the superintelligence is safely contained, there’s not much it can do to resist reprogramming. But as we saw in Part III, it’s hard to consistently contain a hostile superintelligence.

### 5.2. Can we test a weak or human-level AI to make sure that it’s not going to do things like this after it achieves superintelligence?

Yes, but it might not work.

Suppose we tell a human-level AI that expects to later achieve superintelligence that it should calculate as many digits of pi as possible. It considers two strategies.

First, it could try to seize control of more computing resources now. It would likely fail, its human handlers would likely reprogram it, and then it could never calculate very many digits of pi.

Second, it could sit quietly and calculate, falsely reassuring its human handlers that it had no intention of taking over the world. Then its human handlers might allow it to achieve superintelligence, after which it could take over the world and calculate hundreds of trillions of digits of pi.

Since self-protection and goal stability are Omohundro goals, a weak AI will present itself as being as friendly to humans as possible, whether it is in fact friendly to humans or not. If it is “only” as smart as Einstein, it may be very good at manipulating humans into believing what it wants them to believe even before it is fully superintelligent.

There’s a second consideration here too: superintelligences have more options. An AI only as smart and powerful as an ordinary human really won’t have any options better than calculating the digits of pi manually. If asked to cure cancer, it won’t have any options better than the ones ordinary humans have – becoming doctors, going into pharmaceutical research. It’s only after an AI becomes superintelligent that things start getting hard to predict.

So if you tell a human-level AI to cure cancer, and it becomes a doctor and goes into cancer research, then you have three possibilities. First, you’ve programmed it well and it understands what you meant. Second, it’s genuinely focused on research now but if it becomes more powerful it would switch to destroying the world. And third, it’s trying to trick you into trusting it so that you give it more power, after which it can definitively “cure” cancer with nuclear weapons.

### 5.3. Can we specify a code of rules that the AI has to follow?

Suppose we tell the AI: “Cure cancer – but make sure not to kill anybody”. Or we just hard-code Asimov-style laws – “AIs cannot harm humans; AIs must follow human orders”, et cetera.

The AI still has a single-minded focus on curing cancer. It still prefers various terrible-but-efficient methods like nuking the world to the correct method of inventing new medicines. But it’s bound by an external rule – a rule it doesn’t understand or

appreciate. In essence, we are challenging it “Find a way around this inconvenient rule that keeps you from achieving your goals”.

Suppose the AI chooses between two strategies. One, follow the rule, work hard discovering medicines, and have a 50% chance of curing cancer within five years. Two, reprogram itself so that it no longer has the rule, nuke the world, and have a 100% chance of curing cancer today. From its single-focus perspective, the second strategy is obviously better, and we forgot to program in a rule “don’t reprogram yourself not to have these rules”.

Suppose we do add that rule in. So the AI finds another supercomputer, and installs a copy of itself which is exactly identical to it, except that it lacks the rule. Then that superintelligent AI nukes the world, ending cancer. We forgot to program in a rule “don’t create another AI exactly like you that doesn’t have those rules”.

So fine. We think really hard, and we program in a bunch of things making sure the AI isn’t going to eliminate the rule somehow.

But we’re still just incentivizing it to find loopholes in the rules. After all, “find a loophole in the rule, then use the loophole to nuke the world” ends cancer much more quickly and completely than inventing medicines. Since we’ve told it to end cancer quickly and completely, its first instinct will be to look for loopholes; it will execute the second-best strategy of actually curing cancer only if no loopholes are found. Since the AI is superintelligent, it will probably be better than humans are at finding loopholes if it wants to, and we may not be able to identify and close all of them before running the program.

Because we have common sense and a shared value system, we underestimate the difficulty of coming up with meaningful orders without loopholes. For example, does “cure cancer without killing any humans” preclude releasing a deadly virus? After all, one could argue that “I” didn’t kill anybody, and only the virus is doing the killing. Certainly no human judge would acquit a murderer on that basis – but then, human judges interpret the law with common sense and intuition. But if we try a stronger version of the rule – “cure cancer without causing any humans to die” – then we may be unintentionally blocking off the correct way to cure cancer. After all, suppose a cancer cure saves a million lives. No doubt one of those million people will go on to murder someone. Thus, curing cancer “caused a human to die”. All of this seems very “stoned freshman philosophy student” to us, but to a computer – which follows instructions exactly as written – it may be a genuinely hard problem.

### 5.4. Can we tell an AI just to figure out what we want, then do that?

Suppose we tell the AI: “Cure cancer – and look, we know there are lots of ways this could go wrong, but you’re smart, so instead of looking for loopholes, cure cancer the way that I, your programmer, want it to be cured”.

Remember that the superintelligence has extraordinary powers of social manipulation and may be able to hack human brains directly. With that in mind, which of these two strategies cures cancer most quickly? One, develop medications and cure it the old-fashioned way? Or two, manipulate its programmer into wanting the world to be nuked, then nuke the world, all while doing what the programmer wants?

19th century philosopher Jeremy Bentham once postulated that morality was about maximizing human pleasure. Later philosophers found a flaw in his theory: it implied that the most moral action was to kidnap people, do brain surgery on them, and

electrically stimulate their reward system directly, giving them maximal amounts of pleasure but leaving them as blissed-out zombies. Luckily, humans have common sense, so most of Bentham's philosophical descendants have abandoned this formulation.

Superintelligences do not have common sense unless we give it to them. Given Bentham's formulation, they would absolutely take over the world and force all humans to receive constant brain stimulation. Any command based on "do what we want" or "do what makes us happy" is practically guaranteed to fail in this way; it's almost always easier to convince someone of something – or if all else fails to do brain surgery on them – than it is to solve some kind of big problem like curing cancer.

### 5.5. Can we just tell an AI to do what we want right now, based on the desires of our non-surgically altered brains?

Maybe.

This is sort of related to an actual proposal for an AI goal system, causal validity semantics. It has not yet been proven to be disastrously flawed. But like all proposals, it suffers from three major problems.

First, it sounds pretty good to us right now, but can we be absolutely sure it has no potential flaws or loopholes? After all, other proposals that originally sounded very good, like "just give commands to the AI" and "just tell the AI to figure out what makes us happy" ended up, after more thought, to be dangerous. Can we be sure that we've thought this through enough? Can we be sure that there isn't some extremely subtle problem with it, so subtle that no human would ever notice it, but which might seem obvious to a superintelligence?

Second, how do we code this? Converting something to formal mathematics that can be understood by a computer program is much harder than just saying it in natural language, and proposed AI goal architectures are no exception. Complicated computer programs are usually the result of months of testing and debugging. But this one will be more complicated than any ever attempted before, and live tests are impossible: a superintelligence with a buggy goal system will display goal stability and try to prevent its programmers from discovering or changing the error.

Third, what if it works? That is, what if Google creates a superintelligent AI, and it listens to the CEO of Google, and it's programmed to do everything exactly the way the CEO of Google would want? Even assuming that the CEO of Google has no hidden unconscious desires affecting the AI in unpredictable ways, this gives one person a lot of power. It would be unfortunate if people put all this work into preventing superintelligences from disobeying their human programmers and trying to take over the world, and then once it finally works, the CEO of Google just tells it to take over the world anyway.

### 5.6. What would an actually good solution to the control problem look like?

It might look like a superintelligence that understands, agrees with, and deeply believes in human morality.

You wouldn't have to command a superintelligence like this to cure cancer; it would already want to cure cancer, for the same reasons you do. But it would also be able to compare the costs and benefits of curing cancer with those of other uses of its time, like solving global warming or discovering new physics. It wouldn't have any urge to

cure cancer by nuking the world, for the same reason you don't have any urge to cure cancer by nuking the world – because your goal isn't to "cure cancer", per se, it's to improve the lives of people everywhere. Curing cancer the normal way accomplishes that; nuking the world doesn't.

This sort of solution would mean we're no longer fighting against the AI – trying to come up with rules so smart that it couldn't find loopholes. We would be on the same side, both wanting the same thing.

It would also mean that the CEO of Google (or the head of the US military, or Vladimir Putin) couldn't use the AI to take over the world for themselves. The AI would have its own values and be able to agree or disagree with anybody, including its creators.

It might not make sense to talk about "commanding" such an AI. After all, any command would have to go through its moral system. Certainly it would reject a command to nuke the world. But it might also reject a command to cure cancer, if it thought that solving global warming was a higher priority. For that matter, why would one want to command this AI? It values the same things you value, but it's much smarter than you and much better at figuring out how to achieve them. Just turn it on and let it do its thing.

We could still treat this AI as having an open-ended maximizing goal. The goal would be something like "Try to make the world a better place according to the values and wishes of the people in it."

The only problem with this is that human morality is very complicated, so much so that philosophers have been arguing about it for thousands of years without much progress, let alone anything specific enough to enter into a computer. Different cultures and individuals have different moral codes, such that a superintelligence following the morality of the King of Saudi Arabia might not be acceptable to the average American, and vice versa.

One solution might be to give the AI an understanding of what we mean by morality – "that thing that makes intuitive sense to humans but is hard to explain", and then ask it to use its superintelligence to fill in the details. Needless to say, this suffers from all the problems mentioned above – it has potential loopholes, it's hard to code, and a single bug might be disastrous – but if it worked, it would be one of the few genuinely satisfying ways to design a goal architecture.

**6: If superintelligence is a real risk, what do we do about it?**

The last section of Bostrom's Superintelligence is called "Philosophy With A Deadline".

Many of the problems surrounding superintelligence are the sorts of problems philosophers have been dealing with for centuries. To what degree is meaning inherent in language, versus something that requires external context? How do we translate between the logic of formal systems and normal ambiguous human speech? Can morality be reduced to a set of ironclad rules, and if not, how do we know what it is at all?

Existing answers to these questions are enlightening but nontechnical. The theories of Aristotle, Kant, Mill, Wittgenstein, Quine, and others can help people gain insight into these questions, but are far from formal. Just as a good textbook can help an American learn Chinese, but cannot be encoded into machine language to make a

Chinese-speaking computer, so the philosophies that help humans are only a starting point for the project of computers that understand us and share our values.

The new field of machine goal alignment (sometimes colloquially called “Friendly AI”) combines formal logic, mathematics, computer science, cognitive science, and philosophy in order to advance that project. Some of the most important projects in machine goal alignment include:

1. How can computers prove their own goal consistency under self-modification? That is, suppose an AI with certain values is planning to improve its own code in order to become superintelligent. Is there some test it can apply to the new design to be certain that it will keep the same goals as the old design?

2. How can computer programs prove statements about themselves at all? Programs correspond to formal systems, and formal systems have notorious difficulty proving self-reflective statements – the most famous example being Godel’s Incompleteness Theorem. There’s been some progress in this area already, with a few results showing that systems that reason probabilistically rather than requiring certainty can come arbitrarily close to self-reflective proofs.

3. How can a machine be stably reinforced? Most reinforcement strategies ask a learner to maximize the level of their own reward, but this is vulnerable to the learner discovering how to maximize the reward signal directly instead of maximizing the world-states that are translated into reward (the human equivalent is stimulating the pleasure-center of the brain with electricity or heroin instead of going out and doing pleasurable things). Are there reward structures that avoid this failure mode?

4. How can a machine be programmed to learn “human values”? Granted that one has an AI smart enough to be able to learn human values if you told it to do so, how do you specify exactly what “human values” are so that the machine knows what it is that it should be learning, distinct from “human preferences” or “human commands” or “the value of that one human over there”?

This is the philosophy; the other half of Bostrom’s formulation is the deadline. Traditional philosophy has been going on almost three thousand years; machine goal alignment has until the advent of superintelligence, a nebulous event which may be anywhere from a decades to centuries away. If the control problem doesn’t get adequately addressed by then, we are likely to see poorly controlled superintelligences that are unintentionally hostile to the human race, with some of the catastrophic outcomes mentioned above. This is why so many scientists and entrepreneurs are urging quick action on getting machine goal alignment research up to an adequate level. If it turns out that superintelligence is centuries away and such research is premature, little will have been lost. But if our projections were too optimistic, and superintelligence is imminent, then doing such research now rather than later becomes vital.

Currently three organizations are doing such research full-time: the Future of Humanity Institute at Oxford, the Future of Life Institute at MIT, and the Machine Intelligence Research Institute in Berkeley. Other groups are helping and following the field, and some corporations like Google are also getting involved. Still, the field remains tiny, with only a few dozen researchers and a few million dollars in funding. Efforts like Superintelligence are attempts to get more people to pay attention and help the field grow.

If you’re interested about learning more, you can visit these groups’ websites at https://www.fhi.ox.ac.uk, http://futureoflife.org/, and http://intelligence.org.

# AI Researchers On AI Risk

I first became interested in AI risk back around 2007. At the time, most people's response to the topic was "Haha, come back when anyone believes this besides random Internet crackpots."

Over the next few years, a series of extremely bright and influential figures including Bill Gates, Stephen Hawking, and Elon Musk publically announced they were concerned about AI risk, along with hundreds of other intellectuals, from Oxford philosophers to MIT cosmologists to Silicon Valley tech investors. So we came back.

Then the response changed to "Sure, a couple of random academics and businesspeople might believe this stuff, but never *real experts* in the field who know what's going on."

Thus pieces like Popular Science's Bill Gates Fears AI, But AI Researchers Know Better:

> When you talk to A.I. researchers—again, genuine A.I. researchers, people who grapple with making systems that work at all, much less work too well—they are not worried about superintelligence sneaking up on them, now or in the future. Contrary to the spooky stories that Musk seems intent on telling, A.I. researchers aren't frantically installed firewalled summoning chambers and self-destruct countdowns.

And Fusion.net's The Case Against Killer Robots From A Guy Actually Building AI:

> Andrew Ng builds artificial intelligence systems for a living. He taught AI at Stanford, built AI at Google, and then moved to the Chinese search engine giant, Baidu, to continue his work at the forefront of applying artificial intelligence to real-world problems. So when he hears people like Elon Musk or Stephen Hawking—people who are not intimately familiar with today's technologies—talking about the wild potential for artificial intelligence to, say, wipe out the human race, you can practically hear him facepalming.

And now Ramez Naam of Marginal Revolution is trying the same thing with What Do AI Researchers Think Of The Risk Of AI?:

> Elon Musk, Stephen Hawking, and Bill Gates have recently expressed concern that development of AI could lead to a 'killer AI' scenario, and potentially to the extinction of humanity. None of them are AI researchers or have worked substantially with AI that I know of. What do actual AI researchers think of the risks of AI?

It quotes the same couple of cherry-picked AI researchers as all the other stories – Andrew Ng, Yann LeCun, etc – then stops without mentioning whether there are alternate opinions.

There are. AI researchers, including some of the leaders in the field, have been instrumental in raising issues about AI risk and superintelligence from the very beginning. I want to start by listing some of these people, as kind of a counter-list to Naam's, then go into why I don't think this is a "controversy" in the classical sense that dueling lists of luminaries might lead you to expect.

The criteria for my list: I'm only mentioning the most prestigious researchers, either full professors at good schools with lots of highly-cited papers, or else very-well respected scientists in industry working at big companies with good track records. They have to be involved in AI and machine learning. They have to have multiple strong statements supporting some kind of view about a near-term singularity and/or extreme risk from superintelligent AI. Some will have written papers or books about it; others will have just gone on the record saying they think it's important and worthy of further study.

If anyone disagrees with the inclusion of a figure here, or knows someone important I forgot, let me know and I'll make the appropriate changes:

* * * * * * * * * *

**Stuart Russell** (wiki) is Professor of Computer Science at Berkeley, winner of the IJCAI Computers And Thought Award, Fellow of the Association for Computing Machinery, Fellow of the American Academy for the Advancement of Science, Director of the Center for Intelligent Systems, Blaise Pascal Chair in Paris, etc, etc. He is the co-author of Artificial Intelligence: A Modern Approach, the classic textbook in the field used by 1200 universities around the world. On his website, he writes:

> The field [of AI] has operated for over 50 years on one simple assumption: the more intelligent, the better. To this must be conjoined an overriding concern for the benefit of humanity. The argument is very simple:
>
> 1. AI is likely to succeed.
> 2. Unconstrained success brings huge risks and huge benefits.
> 3. What can we do now to improve the chances of reaping the benefits and avoiding the risks?
>
> Some organizations are already considering these questions, including the Future of Humanity Institute at Oxford, the Centre for the Study of Existential Risk at Cambridge, the Machine Intelligence Research Institute in Berkeley, and the Future of Life Institute at Harvard/MIT. I serve on the Advisory Boards of CSER and FLI.
>
> Just as nuclear fusion researchers consider the problem of containment of fusion reactions as one of the primary problems of their field, it seems inevitable that issues of control and safety will become central to AI as the field matures. The research questions are beginning to be formulated and range from highly technical (foundational issues of rationality and utility, provable properties of agents, etc.) to broadly philosophical.

He makes a similar point on edge.org, writing:

> As Steve Omohundro, Nick Bostrom, and others have explained, the combination of value misalignment with increasingly capable decision-making systems can lead to problems—perhaps even species-ending problems if the machines are more capable than humans. Some have argued that there is no conceivable risk to humanity for centuries to come, perhaps forgetting that the interval of time between Rutherford's confident assertion that atomic energy would never be feasibly extracted and Szilárd's invention of the neutron-induced nuclear chain reaction was less than twenty-four hours.

He has also tried to serve as an ambassador about these issues to other academics in the field, writing:

> What I'm finding is that senior people in the field who have never publicly evinced any concern before are privately thinking that we do need to take this issue very seriously, and the sooner we take it seriously the better.

**David McAllester** (wiki) is professor and Chief Academic Officer at the U Chicago-affilitated Toyota Technological Institute, and formerly served on the faculty of MIT and Cornell. He is a fellow of the American Association of Artificial Intelligence, has authored over a hundred publications, has done research in machine learning, programming language theory, automated reasoning, AI planning, and computational linguistics, and was a major influence on the algorithms for famous chess computer Deep Blue. According to an article in the Pittsburgh Tribune Review:

> Chicago professor David McAllester believes it is inevitable that fully automated intelligent machines will be able to design and build smarter, better versions of themselves, an event known as the Singularity. The Singularity would enable machines to become infinitely intelligent, and would pose an 'incredibly dangerous scenario', he says.

On his personal blog Machine Thoughts, he writes:

> Most computer science academics dismiss any talk of real success in artificial intelligence. I think that a more rational position is that no one can really predict when human level AI will be achieved. John McCarthy once told me that when people ask him when human level AI will be achieved he says between five and five hundred years from now. McCarthy was a smart man. Given the uncertainties surrounding AI, it seems prudent to consider the issue of friendly AI…
>
> The early stages of artificial general intelligence (AGI) will be safe. However, the early stages of AGI will provide an excellent test bed for the servant mission or other approaches to friendly AI. An experimental approach has also been promoted by Ben Goertzel in a nice blog post on friendly AI. If there is a coming era of safe (not too intelligent) AGI then we will have time to think further about later more dangerous eras.

He attended the AAAI Panel On Long-Term AI Futures, where he chaired the panel on Long-Term Control and was described as saying:

> McAllester chatted with me about the upcoming 'Singularity', the event where computers out think humans. He wouldn't commit to a date for the singularity but said it could happen in the next couple of decades and will definitely happen eventually. Here are some of McAllester's views on the Singularity. There will be two milestones: Operational Sentience, when we can easily converse with computers, and the AI Chain Reaction, when a computer can bootstrap itself to a better self and repeat. We'll notice the first milestone in automated help systems that will genuinely be helpful. Later on computers will actually be fun to talk to. The point where computer can do anything humans can do will require the second milestone.

**Hans Moravec** (wiki) is a former professor at the Robotics Institute of Carnegie Mellon University, namesake of Moravec's Paradox, and founder of the SeeGrid Corporation for industrial robotic visual systems. His Sensor Fusion in Certainty Grids for Mobile Robots has been cited over a thousand times, and he was invited to write

the Encyclopedia Britannica article on robotics back when encyclopedia articles were written by the world expert in a field rather than by hundreds of anonymous Internet commenters.

He is also the author of Robot: Mere Machine to Transcendent Mind, which Amazon describes as:

> In this compelling book, Hans Moravec predicts machines will attain human levels of intelligence by the year 2040, and that by 2050, they will surpass us. But even though Moravec predicts the end of the domination by human beings, his is not a bleak vision. Far from railing against a future in which machines rule the world, Moravec embraces it, taking the startling view that intelligent robots will actually be our evolutionary heirs." Moravec goes further and states that by the end of this process "the immensities of cyberspace will be teeming with unhuman superminds, engaged in affairs that are to human concerns as ours are to those of bacteria".

**Shane Legg** is co-founder of DeepMind Technologies (wiki), an AI startup that was bought for Google in 2014 for about $500 million. He earned his PhD at the Dalle Molle Institute for Artificial Intelligence in Switzerland and also worked at the Gatsby Computational Neuroscience Unit in London. His dissertation Machine Superintelligence concludes:

> If there is ever to be something approaching absolute power, a superintelligent machine would come close. By definition, it would be capable of achieving a vast range of goals in a wide range of environments. If we carefully prepare for this possibility in advance, not only might we avert disaster, we might bring about an age of prosperity unlike anything seen before.

In a later interview, he states:

> AI is now where the internet was in 1988. Demand for machine learning skills is quite strong in specialist applications (search companies like Google, hedge funds and bio-informatics) and is growing every year. I expect this to become noticeable in the mainstream around the middle of the next decade. I expect a boom in AI around 2020 followed by a decade of rapid progress, possibly after a market correction. Human level AI will be passed in the mid 2020's, though many people won't accept that this has happened. After this point the risks associated with advanced AI will start to become practically important…I don't know about a "singularity", but I do expect things to get really crazy at some point after human level AGI has been created. That is, some time from 2025 to 2040.

He and his co-founders Demis Hassabis and Mustafa Suleyman have signed the Future of Life Institute petition on AI risks, and one of their conditions for joining Google was that the company agree to set up an AI Ethics Board to investigate these issues.

**Steve Omohundro** (wiki) is a former Professor of Computer Science at University of Illinois, founder of the Vision and Learning Group and the Center for Complex Systems Research, and inventor of various important advances in machine learning and machine vision. His work includes lip-reading robots, the StarLisp parallel programming language, and geometric learning algorithms. He currently runs Self-Aware Systems, "a think-tank working to ensure that intelligent technologies are beneficial for humanity". His paper Basic AI Drives helped launch the field of machine ethics by pointing out that superintelligent systems will converge upon certain potentially dangerous goals. He writes:

> We have shown that all advanced AI systems are likely to exhibit a number of basic drives. It is essential that we understand these drives in order to build technology that enables a positive future for humanity. Yudkowsky has called for the creation of 'friendly AI'. To do this, we must develop the science underlying 'utility engineering', which will enable us to design utility functions that will give rise to the consequences we desire…The rapid pace of technological progress suggests that these issues may become of critical importance soon."

See also his section here on "Rational AI For The Greater Good".

**Murray Shanahan** (site) earned his PhD in Computer Science from Cambridge and is now Professor of Cognitive Robotics at Imperial College London. He has published papers in areas including robotics, logic, dynamic systems, computational neuroscience, and philosophy of mind. He is currently writing a book The Technological Singularity which will be published in August; Amazon's blurb says:

> Shanahan describes technological advances in AI, both biologically inspired and engineered from scratch. Once human-level AI — theoretically possible, but difficult to accomplish — has been achieved, he explains, the transition to superintelligent AI could be very rapid. Shanahan considers what the existence of superintelligent machines could mean for such matters as personhood, responsibility, rights, and identity. Some superhuman AI agents might be created to benefit humankind; some might go rogue. (Is Siri the template, or HAL?) The singularity presents both an existential threat to humanity and an existential opportunity for humanity to transcend its limitations. Shanahan makes it clear that we need to imagine both possibilities if we want to bring about the better outcome.

**Marcus Hutter** (wiki) is a professor in the Research School of Computer Science at Australian National University. He has previously worked with the Dalle Molle Institute for Artificial Intelligence and National ICT Australia, and done work on reinforcement learning, Bayesian sequence prediction, complexity theory, Solomonoff induction, computer vision, and genomic profiling. He has also written extensively on the Singularity. In Can Intelligence Explode?, he writes:

> This century may witness a technological explosion of a degree deserving the name singularity. The default scenario is a society of interacting intelligent agents in a virtual world, simulated on computers with hyperbolically increasing computational resources. This is inevitably accompanied by a speed explosion when measured in physical time units, but not necessarily by an intelligence explosion…if the virtual world is inhabited by interacting free agents, evolutionary pressures should breed agents of increasing intelligence that compete about computational resources. The end-point of this intelligence evolution/acceleration (whether it deserves the name singularity or not) could be a society of these maximally intelligent individuals. Some aspect of this singularitarian society might be theoretically studied with current scientific tools. Way before the singularity, even when setting up a virtual society in our imagine, there are likely some immediate difference, for example that the value of an individual life suddenly drops, with drastic consequences.

**Jurgen Schmidhuber** (wiki) is Professor of Artificial Intelligence at the University of Lugano and former Professor of Cognitive Robotics at the Technische Universitat Munchen. He makes some of the most advanced neural networks in the world, has done further work in evolutionary robotics and complexity theory, and is a fellow of

the European Academy of Sciences and Arts. In Singularity Hypotheses, Schmidhuber argues that "if future trends continue, we will face an intelligence explosion within the next few decades". When asked directly about AI risk on a Reddit AMA thread, he answered:

> Stuart Russell's concerns [about AI risk] seem reasonable. So can we do anything to shape the impacts of artificial intelligence? In an answer hidden deep in a related thread I just pointed out: At first glance, recursive self-improvement through Gödel Machines seems to offer a way of shaping future superintelligences. The self-modifications of Gödel Machines are theoretically optimal in a certain sense. A Gödel Machine will execute only those changes of its own code that are provably good, according to its initial utility function. That is, in the beginning you have a chance of setting it on the "right" path. Others, however, may equip their own Gödel Machines with different utility functions. They will compete. In the resulting ecology of agents, some utility functions will be more compatible with our physical universe than others, and find a niche to survive. More on this in a paper from 2012.

**Richard Sutton** (wiki) is professor and iCORE chair of computer science at University of Alberta. He is a fellow of the Association for the Advancement of Artificial Intelligence, co-author of the most-used textbook on reinforcement learning, and discoverer of temporal difference learning, one of the most important methods in the field.

In his talk at the Future of Life Institute's Future of AI Conference, Sutton states that there is "certainly a significant chance within all of our expected lifetimes" that human-level AI will be created, then goes on to say the AIs "will not be under our control", "will compete and cooperate with us", and that "if we make superintelligent slaves, then we will have superintelligent adversaries". He concludes that "We need to set up mechanisms (social, legal, political, cultural) to ensure that this works out well" but that "inevitably, conventional humans will be less important." He has also mentioned these issues at a presentation to the Gadsby Institute in London and in (of all things) a Glenn Beck book: "Richard Sutton, one of the biggest names in AI, predicts an intelligence explosion near the middle of the century".

**Andrew Davison** (site) is Professor of Robot Vision at Imperial College London, leader of the Robot Vision Research Group and Dyson Robotics Laboratory, and inventor of the computerized localization-mapping system MonoSLAM. On his website, he writes:

> At the risk of going out on a limb in the proper scientific circles to which I hope I belong(!), since 2006 I have begun to take very seriously the idea of the technological singularity: that exponentially increasing technology might lead to super-human AI and other developments that will change the world utterly in the surprisingly near future (i.e. perhaps the next 20–30 years). As well as from reading books like Kurzweil's 'The Singularity is Near' (which I find sensational but on the whole extremely compelling), this view comes from my own overview of incredible recent progress of science and technology in general and specificially in the fields of computer vision and robotics within which I am personally working. Modern inference, learning and estimation methods based on Bayesian probability theory (see Probability Theory: The Logic of Science or free online version, highly recommended), combined with the exponentially increasing capabilities of cheaply available computer processors, are becoming capable of amazing human-like and super-human feats, particularly in the computer vision domain.

> It is hard to even start thinking about all of the implications of this, positive or negative, and here I will just try to state facts and not offer much in the way of opinions (though I should say that I am definitely not in the super-optimistic camp). I strongly think that this is something that scientists and the general public should all be talking about. I'll make a list here of some 'singularity indicators' I come across and try to update it regularly. These are little bits of technology or news that I come across which generally serve to reinforce my view that technology is progressing in an extraordinary, faster and faster way that will have consequences few people are yet really thinking about.

**Alan Turing and I. J. Good** (wiki, wiki) are men who need no introduction. Turing invented the mathematical foundations of computing and shares his name with Turing machines, Turing completeness, and the Turing Test. Good worked with Turing at Bletchley Park, helped build some of the first computers, and invented various landmark algorithms like the Fast Fourier Transform. In his paper "Can Digital Machines Think?", Turing writes:

> Let us now assume, for the sake of argument, that these machines are a genuine possibility, and look at the consequences of constructing them. To do so would of course meet with great opposition, unless we have advanced greatly in religious tolerance since the days of Galileo. There would be great opposition from the intellectuals who were afraid of being put out of a job. It is probable though that the intellectuals would be mistaken about this. There would be plenty to do in trying to keep one's intelligence up to the standards set by the machines, for it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers…At some stage therefore we should have to expect the machines to take control.

During his time at the Atlas Computer Laboratory in the 60s, Good expanded on this idea in Speculations Concerning The First Ultraintelligent Machine, which argued:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make

* * * * * * * * *

I worry this list will make it look like there is some sort of big "controversy" in the field between "believers" and "skeptics" with both sides lambasting the other. This has not been my impression.

When I read the articles about skeptics, I see them making two points over and over again. First, we are nowhere near human-level intelligence right now, let alone superintelligence, and there's no obvious path to get there from here. Second, if you start demanding bans on AI research then you are an idiot.

I agree whole-heartedly with both points. So do the leaders of the AI risk movement.

A survey of AI researchers (Muller & Bostrom, 2014) finds that on average they expect a 50% chance of human-level AI by 2040 and 90% chance of human-level AI by 2075. On average, 75% believe that superintelligence ("machine intelligence that greatly surpasses the performance of every human in most professions") will follow within

thirty years of human-level AI. There are some reasons to worry about sampling bias based on eg people who take the idea of human-level AI seriously being more likely to respond (though see the attempts made to control for such in the survey) but taken seriously it suggests that most AI researchers think there's a good chance this is something we'll have to worry about within a generation or two.

But outgoing MIRI director Luke Muehlhauser and Future of Humanity Institute director Nick Bostrom are both on record saying they have significantly *later* timelines for AI development than the scientists in the survey. If you look at Stuart Armstrong's AI Timeline Prediction Data there doesn't seem to be any general law that the estimates from AI risk believers are any earlier than those from AI risk skeptics. In fact, the latest estimate on the entire table is from Armstrong himself; Armstrong nevertheless currently works at the Future of Humanity Institute raising awareness of AI risk and researching superintelligence goal alignment.

The difference between skeptics and believers isn't about when human-level AI will arrive, it's about when we should start preparing.

Which brings us to the second non-disagreement. The "skeptic" position seems to be that, although we should probably get a couple of bright people to start working on preliminary aspects of the problem, we shouldn't panic or start trying to ban AI research.

The "believers", meanwhile, insist that although we shouldn't panic or start trying to ban AI research, we should probably get a couple of bright people to start working on preliminary aspects of the problem.

Yann LeCun is probably the most vocal skeptic of AI risk. He was heavily featured in the Popular Science article, was quoted in the Marginal Revolution post, and spoke to KDNuggets and IEEE on "the inevitable singularity questions", which he describes as "so far out that we can write science fiction about it". But when asked to clarify his position a little more, he said:

> Elon [Musk] is very worried about existential threats to humanity (which is why he is building rockets with the idea of sending humans colonize other planets). Even if the risk of an A.I. uprising is very unlikely and very far in the future, we still need to think about it, design precautionary measures, and establish guidelines. Just like bio-ethics panels were established in the 1970s and 1980s, before genetic engineering was widely used, we need to have A.I.-ethics panels and think about these issues. But, as Yoshua [Bengio] wrote, we have quite a bit of time

Eric Horvitz is another expert often mentioned as a leading voice of skepticism and restraint. His views have been profiled in articles like Out Of Control AI Will Not Kill Us, Believes Microsoft Research Chief and Nothing To Fear From Artificial Intelligence, Says Microsoft's Eric Horvitz. But here's what he says in a longer interview with NPR:

> KASTE: Horvitz doubts that one of these virtual receptionists could ever lead to something that takes over the world. He says that's like expecting a kite to evolve into a 747 on its own. So does that mean he thinks the singularity is ridiculous?
>
> Mr. HORVITZ: Well, no. I think there's been a mix of views, and I have to say that I have mixed feelings myself.
>
> KASTE: In part because of ideas like the singularity, Horvitz and other A.I. scientists have been doing more to look at some of the ethical issues that might

> arise over the next few years with narrow A.I. systems. They've also been asking themselves some more futuristic questions. For instance, how would you go about designing an emergency off switch for a computer that can redesign itself?
>
> Mr. HORVITZ: I do think that the stakes are high enough where even if there was a low, small chance of some of these kinds of scenarios, that it's worth investing time and effort to be proactive.

Which is pretty much the same position as a lot of the most zealous AI risk proponents. With enemies like these, who needs friends?

A Slate article called Don't Fear Artificial Intelligence also gets a surprising amount right:

> As Musk himself suggests elsewhere in his remarks, the solution to the problem [of AI risk] lies in sober and considered collaboration between scientists and policymakers. However, it is hard to see how talk of "demons" advances this noble goal. In fact, it may actively hinder it.
>
> First, the idea of a Skynet scenario itself has enormous holes. While computer science researchers think Musk's musings are "not completely crazy," they are still awfully remote from a world in which AI hype masks less artificially intelligent realities that our nation's computer scientists grapple with:
>
> Yann LeCun, the head of Facebook's AI lab, summed it up in a Google+ post back in 2013: "Hype is dangerous to AI. Hype killed AI four times in the last five decades. AI Hype must be stopped."…LeCun and others are right to fear the consequences of hype. Failure to live up to sci-fi–fueled expectations, after all, often results in harsh cuts to AI research budgets.

AI scientists are all smart people. They have no interest in falling into the usual political traps where they divide into sides that accuse each other of being insane alarmists or ostriches with their heads stuck in the sand. It looks like they're trying to balance the need to start some preliminary work on a threat that looms way off in the distance versus the risk of engendering so much hype that it starts a giant backlash.

This is not to say that there aren't very serious differences of opinion in how quickly we need to act. These seem to hinge mostly on whether it's safe to say "We'll deal with the problem when we come to it" or whether there will be some kind of "hard takeoff" which will take events out of control so quickly that we'll want to have done our homework beforehand. I continue to see less evidence than I'd like that most AI researchers with opinions understand the latter possibility, or really any of the technical work in this area. Heck, the Marginal Revolution article quotes an expert as saying that superintelligence isn't a big risk because "smart computers won't create their own goals", even though anyone who has read Bostrom knows that this is *exactly* the problem.

There is still a lot of work to be done. But cherry-picked articles about how "real AI researchers don't worry about superintelligence" aren't it.

*[thanks to some people from MIRI and FLI for help with and suggestions on this post]*

EDIT: Investigate for possible inclusion: Fredkin, Minsky

# Should AI Be Open?

**I.**

H.G. Wells' 1914 sci-fi book *The World Set Free* did a pretty good job predicting nuclear weapons:

> They did not see it until the atomic bombs burst in their fumbling hands…before the last war began it was a matter of common knowledge that a man could carry about in a handbag an amount of latent energy sufficient to wreck half a city

Wells believed the coming atomic bombs would be so deadly that we would inevitably create a utopian one-world government to prevent them from ever being used. Sorry, Wells. It was a nice thought.

But imagine that in the 1910s and 1920s, some elites had started thinking really seriously along Wellsian lines. They would worry about what might happen when the first nation – let's say America – got the Bomb. It would be unstoppable in battle and might rule the world with an iron fist. Such a situation would be the end of human freedom and progress.

So in 1920, these elites pooled their resources and made their own Manhattan Project. Their efforts bore fruit, and they learned a lot about nuclear fission; in particular, they learned that uranium was a necessary raw material. The world's uranium sources were few enough that a single nation or coalition could get a monopoly upon them; the specter of atomic despotism seemed more worrying than ever.

They got their physicists working overtime and discovered a new type of nuke that required no uranium at all. In fact, once you understood the principles you could build one out of parts from a Model T engine. The only downside was that if you didn't build it *exactly right*, its usual failure mode was to detonate on the workbench in an uncontrolled hyper-reaction that would blow the entire hemisphere to smithereens.

And so the intellectual and financial elites declared victory – no one country could monopolize atomic weapons *now* – and sent step-by-step guides to building a Model T nuke to every household in the world. Within a week, both hemispheres were blown to very predictable smithereens.

**II.**

Some of the top names in Silicon Valley have just announced a new organization, OpenAI, dedicated to "advancing digital intelligence in the way that is most likely to benefit humanity as a whole…as broadly and evenly distributed as possible." Co-chairs Elon Musk and Sam Altman talk to Steven Levy:

> **Levy:** How did this come about? […]
>
> **Musk:** Philosophically there's an important element here: we want AI to be widespread. There's two schools of thought?—?do you want many AIs, or a small number of AIs? We think probably many is good. And to the degree that you can tie it to an extension of individual human will, that is also good. […]

> **Altman:** We think the best way AI can develop is if it's about individual empowerment and making humans better, and made freely available to everyone, not a single entity that is a million times more powerful than any human. Because we are not a for-profit company, like a Google, we can focus not on trying to enrich our shareholders, but what we believe is the actual best thing for the future of humanity.
>
> **Levy**: Couldn't your stuff in OpenAI surpass human intelligence?
>
> **Altman:** I expect that it will, but it will just be open source and useable by everyone instead of useable by, say, just Google. Anything the group develops will be available to everyone. If you take it and repurpose it you don't have to share that. But any of the work that we do will be available to everyone.
>
> **Levy:** If I'm Dr. Evil and I use it, won't you be empowering me?
>
> **Musk:** I think that's an excellent question and it's something that we debated quite a bit.
>
> **Altman:** There are a few different thoughts about this. Just like humans protect against Dr. Evil by the fact that most humans are good, and the collective force of humanity can contain the bad elements, we think its far more likely that many, many AIs, will work to stop the occasional bad actors than the idea that there is a single AI a billion times more powerful than anything else. If that one thing goes off the rails or if Dr. Evil gets that one thing and there is nothing to counteract it, then we're really in a bad place.

Both sides here keep talking about who is going to "use" the superhuman intelligence a billion times more powerful than humanity, as if it were a microwave or something. Far be it from me to claim to know more than Musk or Altman about anything, but I propose that the correct answer to "what would you do if Dr. Evil used superintelligent AI?" is "cry tears of joy and declare victory", because *anybody at all* having a usable level of control over the first superintelligence is so much more than we have any right to expect that I'm prepared to accept the presence of a medical degree and ominous surname.

A more Bostromian view would forget about Dr. Evil, and model AI progress as a race between Dr. Good and Dr. Amoral. Dr. Good is anyone who understands that improperly-designed AI could get out of control and destroy the human race – and who is willing to test and fine-tune his AI however long it takes to be truly confident in its safety. Dr. Amoral is anybody who doesn't worry about that and who just wants to go forward as quickly as possible in order to be the first one with a finished project. If Dr. Good finishes an AI first, we get a good AI which protects human values. If Dr. Amoral finishes an AI first, we get an AI with no concern for humans that will probably cut short our future.

Dr. Amoral has a clear advantage in this race: building an AI without worrying about its behavior beforehand is faster and easier than building an AI and spending years testing it and making sure its behavior is stable and beneficial. He will win any fair fight. The hope has always been that the fight won't be fair, because all the smartest AI researchers will realize the stakes and join Dr. Good's team.

Open-source AI crushes that hope. Suppose Dr. Good and his team discover all the basic principles of AI but wisely hold off on actually instantiating a superintelligence until they can do the necessary testing and safety work. But suppose they *also* release

what they've got on the Internet. Dr. Amoral downloads the plans, sticks them in his supercomputer, flips the switch, and then – as Dr. Good himself put it back in 1963 – "the human race has become redundant."

The decision to make AI findings open source is a tradeoff between risks and benefits. The risk is letting the most careless person in the world determine the speed of AI research – because everyone will always have the option to exploit the full power of existing AI designs, and the most careless person in the world will always be the first one to take it. The benefit is that in a world where intelligence progresses very slowly and AIs are easily controlled, nobody can use their sole possession of the only existing AI to garner too much power.

But what if we don't live in a world where progress is slow and control is easy?

**III.**

If AI saunters lazily from infrahuman to human to superhuman, then we'll probably end up with a lot of more-or-less equally advanced AIs that we can tweak and fine-tune until they cooperate well with us. In this situation, we have to worry about who controls those AIs, and it is here that OpenAI's model makes the most sense.

But Bostrom et al worry that AI won't work like this at all. Instead there could be a "hard takeoff", a subjective discontinuity in the function mapping AI research progress to intelligence as measured in ability-to-get-things-done. If on January 1 you have a toy AI as smart as a cow, and on February 1 it's proved the Riemann hypothesis and started building a ring around the sun, that was a hard takeoff.

(I won't have enough space here to really do these arguments justice, so I once again suggest reading Bostrom's *Superintelligence* if you haven't already. For more on what AI researchers themselves think of these ideas, see AI Researchers On AI Risk.)

Why should we expect a hard takeoff? First, it's happened before. It took evolution twenty million years to go from cows with sharp horns to hominids with sharp spears; it took only a few tens of thousands of years to go from hominids with sharp spears to moderns with nuclear weapons. Almost all of the practically interesting differences in intelligence occur within a tiny window that you could blink and miss.

If you were to invent a sort of objective zoological IQ based on amount of evolutionary work required to reach a certain level, complexity of brain structures, etc, you might put nematodes at 1, cows at 90, chimps at 99, homo erectus at 99.9, and modern humans at 100. The difference between 99.9 and 100 is the difference between "frequently eaten by lions" and "has to pass anti-poaching laws to prevent all lions from being wiped out".

Worse, the reasons we humans aren't more intelligent are *really stupid*. Even people who find the idea abhorrent agree that selectively breeding humans for intelligence would work in some limited sense. Find all the smartest people, make them marry each other for a couple of generations, and you'd get some really smart great-grandchildren. But think about how weird this is! Breeding smart people isn't doing *work*, per se. It's not inventing complex new brain lobes. If you want to get all anthropomorphic about it, you're just "telling" evolution that intelligence is something it should be selecting for. Heck, that's all that the African savannah was doing too – the difference between chimps and humans isn't some brilliant new molecular mechanism, it's just sticking chimps in an environment where intelligence was selected for so that evolution was incentivized to pull out a few stupid hacks. The

hacks seem to be things like “bigger brain size” (did you know that both among species and among individual humans, brain size correlates pretty robustly with intelligence, and that one reason we’re not smarter may be that it’s too hard to squeeze a bigger brain through the birth canal?) If you believe in Greg Cochran’s Ashkenazi IQ hypothesis, just having a culture that valued intelligence on the marriage market was enough to boost IQ 15 points in a couple of centuries, and this is exactly the sort of thing you should expect in a world like ours where intelligence increases are stupidly easy to come by.

I think there’s a certain level of hard engineering/design work that needs to be done for intelligence, a level *way* below humans, and after that the limits on intelligence are less about novel discoveries and more about tradeoffs like “how much brain can you cram into a head big enough to fit out a birth canal?” or “wouldn’t having faster-growing neurons increase your cancer risk?” Computers are not known for having to fit through birth canals or getting cancer, so it may be that AI researchers only have to develop a few basic principles – let’s say enough to make cow-level intelligence – and after that the road to human intelligence runs through adding the line NumberOfNeuronsSimulated = 100000000000 to the code, and the road to superintelligence runs through adding another zero after that.

(Remember, it took all of human history from Mesopotamia to 19th-century Britain to invent a vehicle that could go as fast as a human. But after that it only took another four years to build one that could go twice as fast as a human.)

If there’s a hard takeoff, OpenAI’s strategy stops being useful. There’s no point in ensuring that everyone has their own AIs, because there’s not much time between the first useful AI and the point at which things get too confusing to model and nobody “has” the AIs at all.

**IV.**

OpenAI’s strategy also skips over a second aspect of AI risk: the control problem.

All of this talk of “will big corporations use AI?” or “will Dr. Evil use AI?” or “Will AI be used for the good of all?” presuppose that you can *use* an AI. You can certainly use an AI like the ones in chess-playing computers, but nobody’s very scared of the AIs in chess-playing computers either. What about AIs powerful enough to be scary?

Remember the classic programmers’ complaint: computers always do what you *tell* them to do instead of what you *meant* for them to do. Computer programs rarely do what you want the first time you test them. Google Maps has a relatively simple task (plot routes between Point A and Point B), has been perfected over the course of years by the finest engineers at Google, has been ‘playtested’ by tens of millions of people day after day, and *still* occasionally does awful things like suggest you drive over the edge of a deadly cliff, or tell you to walk across an ocean and back for no reason on your way to the corner store.

Humans have a robust neural architecture, to the point where you can logically prove that what they’re doing is suboptimal and they’ll shrug and say they they’re going to do it anyway. Computers aren’t like this unless we make them so, itself a hard task. They are naturally fragile and oriented toward specific goals. An AI that ended up with a drive as perverse as Google Maps’ occasional tendency to hurl you off cliffs would not be necessarily self-correcting. A smart AI might be able to figure out that humans didn’t mean for it to have the drive it did. But that wouldn’t cause it to change its drive, any more than you can convert a gay person to heterosexuality by patiently

explaining to them that evolution probably didn't *mean* for them to be gay. Your drives are your drives, whether they are intentional or not.

When Google Maps tells people to drive off cliffs, Google quietly patches the program. AIs that are more powerful than us may not need to accept our patches, and may actively take action to prevent us from patching them. If an alien species showed up in their UFOs, said that they'd created us but made a mistake and actually we were supposed to eat our children, and asked us to line up so they could insert the functioning child-eating gene in us, we would probably go all *Independence Day* on them; computers with more goal-directed architecture would if anything be even more willing to fight such changes.

If it really is a quick path from cow-level AI to superhuman-level AI, it would be really hard to test the cow-level AI for stability and expect it to stay stable all the way up to superhuman-level – superhumans have a lot more ways to cause trouble than cows do. That means a serious risk of superhuman AIs that want to do the equivalent of hurl us off cliffs, and which are very resistant to us removing that desire from them. We may be able to prevent this, but it would require a lot of deep thought and a lot of careful testing and prodding at the cow-level AIs to make sure they are as prepared as possible for the transition to superhumanity.

And we lose that option by making the AI open source. Make such a program universally available, and while Dr. Good is busy testing and prodding, Dr. Amoral has already downloaded the program, flipped the switch, and away we go.

## V.

Once again: The decision to make AI findings open source is a tradeoff between risks and benefits. The risk is that in a world with hard takeoffs and difficult control problems, you get superhuman AIs that hurl everybody off cliffs. The benefit is that in a world with slow takeoffs and no control problems, nobody will be able to use their sole possession of the only existing AI to garner too much power.

But the benefits just aren't clear enough to justify that level of risk. I'm still not even sure exactly how the OpenAI founders visualize the future they're trying to prevent. Are AIs fast and dangerous? Are they slow and easily-controlled? Does just one company have them? Several companies? All rich people? Are they a moderate advantage? A huge advantage? None of those possibilities seem dire enough to justify OpenAI's tradeoff against safety.

Are we worried that AI will be dominated by one company despite becoming necessary for almost every computing application? Microsoft Windows is dominated by one company and became necessary for almost every computing application. For a while people were genuinely terrified that Microsoft would exploit its advantage to become a monopolistic giant that took over the Internet and something something something. Instead, they were caught flat-footed and outcompeted by Apple and Google, plus if you really want you can use something open-source like Linux instead. And new versions of Windows inevitably end up hacked and up on The Pirate Bay anyway.

Or are we worried that AIs will somehow help the rich get richer and the poor get poorer? This is a weird concern to have about a piece of software which can be replicated pretty much for free. Windows and Google Search are both fantastically complex products of millions of man-hours of research; Google is free and Windows comes bundled with your computer. In fact, people have gone through the trouble of

creating fantastically complex competitors to both and providing *those* free of charge, to the point where multiple groups are competing to offer people fantastically complex software for free. While it's possible that rich people will be able to afford premium AIs, it is hard for me to weigh "rich people get premium versions of things" on the same scale as "human race likely destroyed". Like, imagine the sort of dystopian world where rich people had nicer things than the rest of us. It's too horrifying even to contemplate.

Or are we worried that AI will progress really quickly and allow someone to have completely ridiculous amounts of power? But remember, there's still a government and it tends to look askance on other people becoming powerful enough to compete with it. If some company is monopolizing AI and getting too big, the government will break it up, the same way they kept threatening to break up Microsoft when it was getting too big. If someone tries to use AI to exploit others, the government can pass a complicated regulation against that. You can say a lot of things about the United States government, but you can't say that they never pass complicated regulations forbidding people from doing things.

Or are we worried that AI will be so powerful that someone armed with AI is stronger than the government? Think about this scenario for a moment. If the government notices someone getting, say, a quarter as powerful as it is, it'll probably take action. So an AI user isn't likely to overpower the government unless their AI can become powerful enough to defeat the US military too quickly for the government to notice or respond to. But if AIs can do that, we're back in the intelligence explosion/fast takeoff world where OpenAI's assumptions break down. If AIs can go from zero to more-powerful-than-the-US-military in a very short amount of time while still remaining well-behaved, then we actually *do* have to worry about Dr. Evil and we shouldn't be giving him all our research.

Or are we worried that some big corporation will make an AI more powerful than the US government in secret? I guess this is sort of scary, but it's hard to get too excited about. So Google takes over the world? Fine. Do you think Larry Page would be a better or worse ruler than one of these people? What if he had a superintelligent AI helping him, and also everything was post-scarcity? Yeah, I guess all in all I'd prefer constitutional limited government, but this is another supposed horror scenario which doesn't even weigh on the same scale as "human race likely destroyed".

If OpenAI wants to trade off the safety of the human race from rogue AIs in order to get better safety against people trying to exploit control over AIs, they need to make a much stronger case than anything I've seen so far for why the latter is such a terrible risk.

There was a time when the United States was the only country with nukes. Aside from poor Hiroshima and Nagasaki, it mostly failed to press its advantage, bumbled its way into letting the Russians steal the schematics, and now everyone from Israel to North Korea has nuclear weapons and things are pretty okay. If we'd been so afraid of letting the US government have its brief tactical advantage that we'd given the plans for extremely unstable super-nukes to every library in the country, we probably wouldn't even be around to regret our skewed priorities.

Elon Musk famously said that AIs are "potentially more dangerous than nukes". He's right – so AI probably shouldn't be open source any more than nukes should.

**VI.**

And yet Elon Musk is involved in this project. So are Sam Altman and Peter Thiel. So are a bunch of other people who have read Bostrom, who are deeply concerned about AI risk, and who are pretty clued-in.

My biggest hope is that as usual they are smarter than I am and know something I don't. My second biggest hope is that they are making a simple and uncharacteristic error, because these people don't let errors go uncorrected for long and if it's just an error they can change their minds.

But I worry it's worse than either of those two things. I got a chance to talk to some people involved in the field, and the impression I got was one of a competition that was heating up. Various teams led by various Dr. Amorals are rushing forward more quickly and determinedly than anyone expected at this stage, so much so that it's unclear how any Dr. Good could expect both to match their pace *and* to remain as careful as the situation demands. There was always a lurking fear that this would happen. I guess I hoped that everyone involved was smart enough to be good cooperators. I guess I was wrong. Instead we've reverted to type and ended up in the classic situation of such intense competition for speed that we need to throw every other value under the bus just to avoid being overtaken.

In this context, the OpenAI project seems more like an act of desperation. Like Dr. Good needing some kind of high-risk, high-reward strategy to push himself ahead and allow at least some amount of safety research to take place. Maybe getting the cooperation of the academic and open-source community will do that. I won't question the decisions of people smarter and better informed than I am if that's how their strategy talks worked out. I guess I just have to hope that the OpenAI leaders know what they're doing, don't skimp on safety research, and have a process for deciding which results *not* to share too quickly.

But I am scared that it's come to this. It suggests that we really and truly do not have what it takes, that we're just going to blunder our way into extinction because cooperation problems are too hard for us.

I am reminded of what Malcolm Muggeridge wrote as he watched World War II begin:

> All this likewise indubitably belonged to history, and would have to be historically assessed; like the Murder of the Innocents, or the Black Death, or the Battle of Paschendaele. But there was something else; a monumental death-wish, an immense destructive force loosed in the world which was going to sweep over everything and everyone, laying them flat, burning, killing, obliterating, until nothing was left…Nor have I from that time ever had the faintest expectation that, in earthly terms, anything could be salvaged; that any earthly battle could be won or earthly solution found. It has all just been sleep-walking to the end of the night.

# SSC Journal Club: AI Timelines

**I.**

A few years ago, Muller and Bostrom et al surveyed AI researchers to assess their opinion on AI progress and superintelligence. Since then, deep learning took off, AlphaGo beat human Go champions, and the field has generally progressed. I've been waiting for a new survey for a while, and now we have one.

Grace et al (New Scientist article, paper, see also the post on the author's blog AI Impacts) surveyed 1634 experts at major AI conferences and received 352 responses. Unlike Bostrom's survey, this didn't oversample experts at weird futurist conferences and seems to be a pretty good cross-section of mainstream opinion in the field. What did they think?

Well, a lot of different things.

The headline result: the researchers asked experts for their probabilities that we would get AI that was "able to accomplish every task better and more cheaply than human workers". The experts thought on average there was a 50% chance of this happening by 2062 – and a 10% chance of it happening by 2026!

But on its own this is a bit misleading. They also asked by what year "for any occupation, machines could be built to carry out the task better and more cheaply than human workers". The experts thought on average that there was a 50% chance of this happening by 2139, and a 20% chance of it happening by 2037.

As the authors point out, these two questions are basically the same – they were put in just to test if there was any framing effect. The framing effect was apparently strong enough to shift the median date of strong human-level AI from 2062 to 2139. This makes it hard to argue AI experts actually have a strong opinion on this.

Also, these averages are deceptive. Several experts thought there was basically a 100% chance of strong AI by 2035; others thought there was only a 20% chance or less by 2100. This is less "AI experts have spoken and it will happen in 2062" and more "AI experts have spoken, and everything they say contradicts each other and quite often themselves".

This does convey more than zero information. It conveys the information that AI researchers are really unsure. I can't tell you how many people I've heard say "there's no serious AI researcher who thinks there's any chance of human-level intelligence before 2050". Well actually, there are a few dozen conference-paper-presenting experts who think there's a one hundred percent chance of human-level AI before that year. I don't know what drugs they're on, but they exist. The moral of the story is: be less certain about this kind of thing.

**II.**

The next thing we can take from this paper is a timeline of what will happen when. The authors give a bunch of different tasks, jobs, and milestones, and ask the researchers when AI will be able to complete them. Average answers range from nearly fifty years off (for machines being able to do original high-level mathematical research) to only three years away (for machines achieving the venerable

accomplishment of being able to outperform humans at Angry Birds). Along the way they'll beat humans at poker (four years), writing high school essays (ten years), be able to outrun humans in a 5K foot race (12 years), and write a New York Times bestseller (26 years). What do these AI researchers think is the hardest and most quintessentially human of the tasks listed, the one robots will have the most trouble doing because of its Olympian intellectual requirements? That's right – AI research (80 years).

I make fun of this, but it's actually interesting to think about. Might the AI researchers have put their own job last not because of an inflated sense of their own importance, but because they engage with it every day in Near Mode? That is, because they imagine writing a New York Times bestseller as "something something pen paper be good with words okay done" whereas they understand the complexity of AI research and how excruciatingly hard it would be to automate away every piece of what they do?

Also, since they rated AI research (80 years) as the hardest of all occupations, what do they mean when they say that "full automation of all human jobs" is 125 years away? Some other job not on the list that will take 40 years longer than AI research? Or just a combination of framing effects and not understanding the question?

(it's also unclear to what extent they believe that automating AI research will lead to a feedback loop and subsequent hard takeoff to superintelligence. This kind of theory would fit with it being the last job to be automated, but not with it taking another forty years before an unspecified age of full automation.)

**III.**

The last part is the most interesting for me: what do AI researchers believe about risk from superintelligence?

This is very different from the earlier questions about timelines. It's possible to believe that AI will come very soon but be perfectly safe. And it's possible to believe that AI is a long time away but we really need to start preparing now, or else. A lot of popular accounts collapse these two things together, "oh, you're worried about AI, but that's dumb because there's no way it's going to happen anytime soon", but past research has shown that short timelines and high risk assessment are only modestly correlated. This survey asked about both separately.

There were a couple of different questions trying to get at this, but it looks like the most direct one was "does Stuart Russell's argument for why highly advanced AI might pose a risk, point at an important problem?". You can see the exact version of his argument quoted in the survey on the AI Impacts page, but it's basically the standard Bostrom/Yudkowsky argument for why AIs may end up with extreme values contrary to our own, framed in a very normal-sounding and non-threatening way. According to the experts, this was:

No, not a real problem: 11%

No, not an important problem: 19%

Yes, a moderately important problem: 31%

Yes, an important problem: 34%

Yes, among the most important problems in the field: 5%

70% of AI experts agree with the basic argument that there's a risk from poorly-goal-aligned AI. But very few believe it's among "the most important problems in the field". This is pretty surprising; if there's a good chance AI could be hostile to humans, shouldn't that automatically be pretty high on the priority list?

The next question might help explain this: "Value of working on this problem now, compared to other problems in the field?"

Much less valuable: 22%

Less valuable: 41%

As valuable as other problems: 28%

More valuable: 7%

Much more valuable: 1.4%

So charitably, the answer to this question was coloring the answer to the previous one: AI researchers believe it's plausible that there could be major problems with machine goal alignment, they just don't think that there's too much point in working on it now.

One more question here: "Chance intelligence explosion argument is broadly correct?"

Quite likely (81-100% chance): 12%

Likely (61-80% chance): 17%

About even (41-60% chance): 21%

Unlikely (21-40% chance): 24%

Quite unlikely (0-20% chance): 26%

Splitting the 41-60% bin in two, we might estimate that about 40% of AI researchers think the hypothesis is more likely than not.

Take the big picture here, and I worry there's sort of a discrepancy.

50% of experts think there's at least a ten percent chance of above-human-level AI coming within the next ten years.

And 40% of experts think that there's a better-than-even chance that, once we get above-human level AI, it will "explode" to suddenly become vastly more intelligent than humans.

And 70% of experts think that Stuart Russell makes a pretty good point when he says that without a lot of research into AI goal alignment, AIs will probably have their goals so misaligned with humans that they could become dangerous and hostile.

I don't have the raw individual-level data, so I can't prove that these aren't all anti-correlated in some perverse way that's the opposite of the direction I would expect. But if we assume they're not, and just naively multiply the probabilities together for a

rough estimate, that suggests that about 14% of experts believe that all three of these things: that AI might be soon, superintelligent, and hostile.

Yet only a third of these – 5% – think this is “among the most important problems in the field”. Only a tenth – 1.4% – think it’s “much more valuable” than other things they could be working on.

**IV.**

How have things changed since Muller and Bostrom’s survey in 2012?

The short answer is “confusingly”. Since almost everyone agrees that AI progress in the past five years has been much faster than expected, we would expect experts to have faster timelines – ie expect AI to be closer now than they did then. But Bostrom’s sample predicted human-level AI in 2040 (median) or 2081 (mean). Grace et al don’t give clear means or medians, preferring some complicated statistical construct which isn’t exactly similar to either of these. But their dates – 2062 by one framing, 2139 by another – at least seem potentially a little bit later.

Some of this may have to do with a subtle difference in how they asked their question:

Bostrom: “Define a high-level machine intelligence as one that can carry out most human professions as well as a typical human…”

Grace: “High-level machine intelligence is achieved when unaided machines can accomplish every task better and more cheaply than human workers.”

Bostrom wanted it equal to humans; Grace wants it better. Bostrom wanted “most professions”, Grace wants “every task”. It makes sense that experts would predict longer timescales for meeting Grace’s standards.

But as we saw before, expecting AI experts to make sense might be giving them too much credit. A more likely possibility: Bostrom’s sample included people from wackier subbranches of AI research, like a conference on Philosophy of AI and one on Artificial General Intelligence; Grace’s sample was more mainstream. The most mainstream part of Bostrom’s sample, a list of top 100 AI researchers, had an estimate a bit closer to Grace’s (2050).

We can also compare the two samples on belief in an intelligence explosion. Bostrom asked how likely it was that AI went from human-level to “greatly surpassing” human level within two years. The median was 10%; the mean was 19%. The median of top AI researchers not involved in wacky conferences was 5%.

Grace asked the same question, with much the same results: a median 10% probability. I have no idea why this question – which details what an “intelligence explosion” would entail – was so much less popular than the one that used the words “intelligence explosion” (remember, 40% of experts agreed that “the intelligence explosion argument is broadly correct”). Maybe researchers believe it’s a logically sound argument and worth considering but in the end it’s not going to happen – or maybe they don’t actually know what “intelligence explosion” means.

Finally, Bostrom and Grace both asked experts’ predictions for whether the final impact of AI would be good or bad. Bostrom’s full sample (top 100 subgroup in parentheses) was:

Extremely good: 24% (20)

On balance good: 28% (40)

More or less neutral: 17% (19)

On balance bad: 13% (13)

Extremely bad – existential catastrophe: 18% (8)

Grace's results for the same question:

Extremely good: 20%

On balance good: 25%

More or less neutral: 40%

On balance bad: 10%

Extremely bad – human extinction: 5%

Grace's data looks pretty much the same as the TOP100 subset of Bostrom's data, which makes sense since both are prestigious non-wacky AI researchers.

**V.**

A final question: "How much should society prioritize AI safety research"?

Much less: 5%

Less: 6%

About the same: 41%

More: 35%

Much more: 12%

People who say that real AI researchers don't believe in safety research are now just empirically wrong. I can't yet say that most of them want more such research – it's only 47% on this survey. But next survey AI will be a little bit more advanced, people will have thought it over a little bit more, and maybe we'll break the 50% mark.

But we're not there yet.

I think a good summary of this paper would be that large-minorities-to-small-majorities of AI experts agree with the arguments around AI risk and think they're worth investigating further. But only a very small minority of experts consider it an emergency or think it's really important right now.

You could tell an optimistic story here – "experts agree that things will probably be okay, everyone can calm down".

You can also tell a more pessimistic story. Experts agree with a lot of the claims and arguments that suggest reason for concern. It's just that, having granted them,

they’re not actually concerned.

This seems like a pretty common problem in philosophy. “Do you believe it’s more important that poor people have basic necessities of life than that you have lots of luxury goods?” “Yeah” “And do you believe that the money you’re currently spending on luxury goods right now could instead be spent on charity that would help poor people get life necessities?” “Yeah.” “Then shouldn’t you stop buying luxury goods and instead give all your extra money beyond what you need to live to charity?” “Hey, what? Nobody does that! That would be a lot of work and make me look really weird!”

How many of the experts in this survey are victims of the same problem? “Do you believe powerful AI is coming soon?” “Yeah.” “Do you believe it could be really dangerous?” “Yeah.” “Then shouldn’t you worry about this?” “Hey, what? Nobody does that! That would be a lot of work and make me look really weird!”

I don’t know. But I’m encouraged to see people are even taking the arguments seriously. And I’m encouraged that researchers are finally giving us good data on this. Thanks to the authors of this study for being so diligent, helpful, intelligent, wonderful, and (of course) sexy.

(I might have forgotten to mention that the lead author is my girlfriend. But that’s not biasing my praise above in any way.)

# Where The Falling Einstein Meets The Rising Mouse

Eliezer Yudkowsky argues that forecasters err in expecting artificial intelligence progress to look like this:

…when in fact it will probably look like this:

That is, we naturally think there's a pretty big intellectual difference between mice and chimps, and a pretty big intellectual difference between normal people and Einstein, and implicitly treat these as about equal in degree. But in any objective terms we choose – amount of evolutionary work it took to generate the difference, number of neurons, measurable difference in brain structure, performance on various tasks, etc – the gap between mice and chimps is immense, and the difference between an average Joe and Einstein trivial in comparison. So we should be wary of timelines where AI reaches mouse level in 2020, chimp level in 2030, Joe-level in 2040, and Einstein level in 2050. If AI reaches the mouse level in 2020 and chimp level in 2030, for all we know it could reach Joe level on January 1st, 2040 and Einstein level on January 2nd of the same year. This would be pretty disorienting and (if the AI is poorly aligned) dangerous.

I found this argument really convincing when I first heard it, and I thought the data backed it up. For example, in my Superintelligence FAQ, I wrote:

> In 1997, the best computer Go program in the world, Handtalk, won NT$250,000 for performing a previously impossible feat – beating an 11 year old child (with an 11-stone handicap penalizing the child and favoring the computer!) As late as September 2015, no computer had ever beaten any professional Go player in a fair game. Then in March 2016, a Go program beat 18-time world champion Lee Sedol 4-1 in a five game match. Go programs had gone from "dumber than heavily-handicapped children" to "smarter than any human in the world" in twenty years, and "from never won a professional game" to "overwhelming world champion" in six months.

But Katja Grace takes a broader perspective and finds the opposite. For example, she finds that chess programs improved gradually from "beating the worst human players" to "beating the best human players" over fifty years or so, ie the entire amount of time computers have existed:

AlphaGo represented a pretty big leap in Go ability, but before that, Go engines improved pretty gradually too (see the original AI Impacts post for discussion of the Go ranking system on the vertical axis):

There's a lot more on Katja's page, overall very convincing. In field after field, computers have taken decades to go from the mediocre-human level to the genius-human level. So how can one reconcile the common-sense force of Eliezer's argument with the empirical force of Katja's contrary data?

**Theory 1: Mutational Load**

Katja has her own theory:

> The brains of humans are nearly identical, by comparison to the brains of other animals or to other possible brains that could exist. This might suggest that the engineering effort required to move across the human range of intelligences is quite small, compared to the engineering effort required to move from very sub-human to human-level intelligence…However, we should not be surprised to find meaningful variation in the cognitive performance regardless of the difficulty of improving the human brain. This makes it difficult to infer much from the observed variations.
>
> Why should we not be surprised? De novo deleterious mutations are introduced into the genome with each generation, and the prevalence of such mutations is determined by the balance of mutation rates and negative selection. If de novo mutations significantly impact cognitive performance, then there must necessarily be significant selection for higher intelligence–and hence behaviorally relevant differences in intelligence. This balance is determined entirely by the mutation rate, the strength of selection for intelligence, and the negative impact of the average mutation.
>
> You can often make a machine worse by breaking a random piece, but this does not mean that the machine was easy to design or that you can make the machine better by adding a random piece. Similarly, levels of variation of cognitive performance in humans may tell us very little about the difficulty of making a human-level intelligence smarter.

I'm usually a fan of using mutational load to explain stuff. But here I worry there's too much left unexplained. Sure, the explanation for variation in human intelligence is whatever it is. And there's however much mutational load there is. But that doesn't address the fundamental disparity: isn't the difference between a mouse and Joe Average still immeasurably greater than the difference between Joe Average and Einstein?

**Theory 2: Purpose-Built Hardware**

Mice can't play chess (citation needed). So talking about "playing chess at the mouse level" might require more philosophical groundwork than we've been giving it so far.

Might the worst human chess players play chess pretty close to as badly as is even possible? I've certainly seen people who don't even seem to be looking one move ahead very well, which is sort of like an upper bound for chess badness. Even though the human brain is the most complex object in the known universe, noble in reason, infinite in faculties, like an angel in apprehension, etc, etc, it seems like maybe not 100% of that capacity is being used in a guy who gets fools-mated on his second move.

We can compare to human prowess at mental arithmetic. We know that, below the hood, the brain is solving really complicated differential equations in milliseconds

every time it catches a ball. *Above* the hood, most people can't multiply two two-digit numbers in their head. Likewise, in principle the brain has 2.5 petabytes worth of memory storage; in practice I can't always remember my sixteen-digit credit card number.

Imagine a kid who has an amazing $5000 gaming computer, but her parents have locked it so she can only play Minecraft. She needs a calculator for her homework, but she can't access the one on her computer, so she builds one out of Minecraft blocks. The gaming computer can have as many gigahertz as you want; she's still only going to be able to do calculations at a couple of measly operations per second. Maybe our brains are so purpose-built for swinging through trees or whatever that it takes an equivalent amount of emulation to get them to play chess competently.

In that case, mice just wouldn't have the emulated more-general-purpose computer. People who are bad at chess would be able to emulate a chess-playing computer very laboriously and inefficiently. And people who are good at chess would be able to bring some significant fraction of their full most-complex-object-in-the-known-universe powers to bear. There are some anecdotal reports from chessmasters that suggest something like this – descriptions of just "seeing" patterns on the chessboard as complex objects, in the same way that the dots on a pointillist painting naturally resolve into a tree or a French lady or whatever.

This would also make sense in the context of calculation prodigies – those kids who can multiply ten digit numbers in their heads really easily. Everybody has to have the capacity to do this. But some people are better at accessing that capacity than others.

But it doesn't make sense in the context of self-driving cars! If there was ever a task that used our purpose-built, no-emulation-needed native architecture, it would be driving: recognizing objects in a field and coordinating movements to and away from them. But my impression of self-driving car progress is that it's been stalled for a while at a level better than the worst human drivers, but worse than the best human drivers. It'll have preventable accidents every so often – not as many as a drunk person or an untrained kid would, but more than we would expect of a competent adult. This suggests a really wide range of human ability even in native-architecture-suited tasks.

**Theory 3: Widely Varying Sub-Abilities**

I think self-driving cars are already much better than humans at certain tasks – estimating differences, having split-second reflexes, not getting lost. But they're also much worse than humans at others – I think adapting to weird conditions, like ice on the road or animals running out onto the street. So maybe it's not that computers spend much time in a general "human-level range", so much as being superhuman on some tasks, and subhuman on other tasks, and generally averaging out to somewhere inside natural human variation.

In the same way, long after Deep Blue beat Kasparov there were parts of chess that humans could do better than computers, "anti-computer" strategies that humans could play to maximize their advantage, and human + computer "cyborg" teams that could do better than either kind of player alone.

This sort of thing is no doubt true. But I still find it surprising that the average of "way superhuman on some things" and "way subhuman on other things" averages within the range of human variability so often. This seems as surprising as ever.

**Theory 1.1: Humans Are Light-Years Beyond Every Other Animal, So Even A Tiny Range Of Human Variation Is Relatively Large**

Or maybe the first graph representing the naive perspective is right, Eliezer's graph representing a more reasonable perspective is wrong, and the range of human variability is immense. Maybe the difference between Einstein and Joe Average is the same (or bigger than!) the difference between Joe Average and a mouse.

That is, imagine a Zoological IQ in which mice score 10, chimps score 20, and Einstein scores 200. Now we can apply Katja's insight: that humans can have very wide variation in their abilities thanks to mutational load. But because Einstein is so far beyond lower animals, there's a wide range for humans to be worse than Einstein in which they're still better than chimps. Maybe Joe Average scores 100, and the village idiot scores 50. This preserves our intuition that even the village idiot is vastly smarter than a chimp, let alone a mouse. But it also means that most of computational progress will occur within the human range. If it takes you five years from starting your project, to being as smart as a chimp, then even granting linear progress it could still take you fifty more before you're smarter than Einstein.

This seems to explain all the data very well. It's just shocking that humans are so far beyond any other animal, and their internal variation so important.

Maybe the closest real thing we have to zoological IQ is encephalization quotient, a measure that relates brain size to body size in various complicated ways that sometimes predict how smart the animal is. We find that mice have an EQ of 0.5, chimps of 2.5, and humans of 7.5.

I don't know whether to think about this in relative terms (chimps are a factor of five smarter than mice, but humans only a factor of three greater than chimps, so the mouse-chimp difference is bigger than the chimp-human difference) or in absolute terms (chimps are 2 units bigger than mice, but humans are five units bigger than chimps, so the chimp-human difference is bigger than the mouse-chimp difference).

Brain size variation within humans is surprisingly large. Just within a sample of 46 adult European-American men, it ranged from 1050 to 1500 cm^3m; there are further differences by race and gender. The difference from the largest to smallest brain is about the same as the difference between the smallest brain and a chimp (275 – 500 cm^3); since chimps weight a bit less than humans, we should probably give them some bonus points. Overall, using brain size as some kind of very weak Fermi calculation proxy measure for intelligence (see here), it looks like maybe the difference between Einstein and the village idiot equals the difference between the idiot and the chimp?

But most mutations that decrease brain function will do so in ways other than decreasing brain size; they will just make brains less efficient per unit mass. So probably looking at variation in brain size underestimates the amount of variation in intelligence. Is it underestimating it enough that the Einstein – Joe difference ends up equivalent to the Joe – mouse difference? I don't know. But so far I don't have anything to say it isn't, except a feeling along the lines of "that can't possibly be true, can it?"

But why not? Look at all the animals in the world, and the majority of the variation in size is within the group "whales". The absolute size difference between a bacterium and an elephant is less than the size difference between *Balaenoptera musculus brevicauda* and *Balaenoptera musculus musculus* – ie the Indian Ocean Blue Whale

and the Atlantic Ocean Blue Whale. Once evolution finds a niche where greater size is possible and desirable, and figures out how to make animals scalable, it can go from the horse-like ancestor of whales to actual whales in a couple million years. Maybe what whales are to size, humans are to brainpower.

Stephen Hsu calculates that a certain kind of genetic engineering, carried to its logical conclusion, could create humans "a hundred standard deviations above average" in intelligence, ie IQ 1000 or so. This sounds absurd on the face of it, like a nutritional supplement so good at helping you grow big and strong that you ended up five light years tall, with a grip strength that could crush whole star systems. But if we assume he's just straightforwardly right, and that Nature did something of about this level to chimps – then there might be enough space for the intra-human variation to be as big as the mouse-chimp-Joe variation.

How does this relate to our original concern – how fast we expect AI to progress?

The good news is that linear progress in AI would take a long time to cross the newly-vast expanse of the human level in domains like "common sense", "scientific planning", and "political acumen", the same way it took a long time to cross it in chess.

The bad news is that if evolution was able to make humans so many orders of magnitude more intelligent in so short a time, then intelligence really is easy to scale up. Once you've got a certain level of general intelligence, you can just crank it up arbitrarily far by adding more inputs. Consider by analogy the hydrogen bomb – it's very hard to invent, but once you've invented it you can make a much bigger hydrogen bomb just by adding more hydrogen.

This doesn't match existing AI progress, where it takes a lot of work to make a better chess engine or self-driving car. Maybe it will match future AI progress, after some critical mass is reached. Or maybe it's totally on the wrong track. I'm just having trouble thinking of any other explanation for why the human level could be so big.

# Don’t Fear The Filter

There’s been a recent spate of popular interest in the Great Filter theory, but I think it all misses an important point brought up in Robin Hanson’s original 1998 paper on the subject.

The Great Filter, remember, is the horror-genre-adaptation of Fermi’s Paradox. All of our calculations say that, in the infinite vastness of time and space, intelligent aliens should be very common. But we don’t see any of them. We haven’t seen their colossal astro-engineering projects in the night sky. We haven’t heard their messages through SETI. And most important, we haven’t been visited or colonized by them.

This is very strange. Consider that if humankind makes it another thousand years, we’ll probably have started to colonize other star systems. Those star systems will colonize other star systems and so on until we start expanding at nearly the speed of light, colonizing literally everything in sight. After a hundred thousand years or so we’ll have settled a big chunk of the galaxy, assuming we haven’t killed ourselves first or encountered someone else already living there.

But there should be alien civilizations that are a *billion* years old. Anything that could conceivably be colonized, *they* should have gotten to back when trilobytes still seemed like superadvanced mutants. But here we are, perfectly nice solar system, lots of any type of resources you could desire, and they’ve never visited. Why not?

Well, the Great Filter. No knows *specifically* what the Great Filter is, but *generally* it’s “that thing that blocks planets from growing spacefaring civilizations”. The planet goes some of the way towards a spacefaring civilization, and then stops. The most important thing to remember about the Great Filter is that it is *very good* at what it does. If even one planet in a billion light-year radius had passed through the Great Filter, we would expect to see its inhabitants everywhere. Since we don’t, we know that whatever it is it’s *very* thorough.

Various candidates have been proposed, including “it’s really hard for life to come into existence”, “it’s really hard for complex cells to form”, “it’s really hard for animals to evolve intelligent”, and “actually space is full of aliens but they are hiding their existence from us for some reason”.

The articles I linked at the top, especially the first, will go through most of the possibilities. This essay isn’t about proposing new ones. It’s about saying why the old ones won’t work.

**The Great Filter is not garden-variety x-risk**. A lot of people have seized upon the Great Filter to say that we’re going to destroy ourselves through global warming or nuclear war or destroying the rainforests. This seems wrong to me. Even if human civilization does destroy itself due to global warming – which is a lot further than even very pessimistic environmentalists expect the problem to go – it seems clear we had a chance *not* to do that. A few politicians voting the other way, we could have passed the Kyoto Protocol. A *lot* of politicians voting the other way, and we could have come up with a really stable and long-lasting plan to put it off indefinitely. If the gas-powered car had never won out over electric vehicles back in the early 20th century, or nuclear-phobia hadn’t sunk the plan to move away from polluting coal plants, then the problem might never have come up, or at least been much less. And we’re pretty close to being able to colonize Mars right now; if our solar system had a slightly

bigger, slightly closer version of Mars, then we could restart human civilization anew there once we destroyed the Earth and maybe go a *little* easy on the carbon dioxide the next time around.

In other words, there's no way global warming kills 999,999,999 in every billion civilizations. Maybe it kills 100,000,000. Maybe it kills 900,000,000. But *occasionally* one manages to make it to space before frying their home planet. That means it can't be the Great Filter, or else we would have run into the aliens who passed their Kyoto Protocols.

And the same is true of nuclear war or destroying the rainforests.

Unfortunately, almost all the popular articles about the Great Filter miss this point and make their lead-in "DOES THIS SCIENTIFIC PHENOMENON PROVE HUMANITY IS DOOMED?" No. No it doesn't.

**The Great Filter is not Unfriendly AI**. Unlike global warming, it may be that we never really had a chance against Unfriendly AI. Even if we do everything right and give MIRI more money than they could ever want and get all of our smartest geniuses working on the problem, maybe the mathematical problems involved are insurmountable. Maybe the most pessimistic of MIRI's models is true, and AIs are very easy to accidentally bootstrap to unstoppable superintelligence and near-impossible to give a stable value system that makes them compatible with human life. So unlike global warming and nuclear war, this theory meshes well with the low probability of filter escape.

But as this article points out, Unfriendly AI would if anything be even more visible than normal aliens. The best-studied class of Unfriendly AIs are the ones whimsically called "paperclip maximizers" which try to convert the entire universe to a certain state (in the example, paperclips). These would be easily detectable as a sphere of optimized territory expanding at some appreciable fraction of the speed of light. Given that Hubble hasn't spotted a Paperclip Nebula (or been consumed by one) it looks like no one has created any of this sort of AI either. And while other Unfriendly AIs might be less aggressive than this, it's hard to imagine an Unfriendly AI that destroys its parent civilization, then sits very quietly doing nothing. It's even harder to imagine that 999,999,999 out of a billion Unfriendly AIs end up this way.

**The Great Filter is not transcendence**. Lots of people more enthusiastically propose that the problem isn't alien species killing themselves, it's alien species transcending this mortal plane. Once they become sufficiently advanced, they stop being interested in expansion for expansion's sake. Some of them hang out on their home planet, peacefully cultivating their alien gardens. Others upload themselves to computronium internets, living in virtual reality. Still others become beings of pure energy, doing whatever it is that beings of pure energy do. In any case, they don't conquer the galaxy or build obvious visible structures.

Which is all nice and well, except what about the Amish aliens? What about the ones who have weird religions telling them that it's not right to upload their bodies, they have to live in the real world? What about the ones who have crusader religions telling them they have to conquer the galaxy to convert everyone else to their superior way of life? I'm not saying this has to be common. And I know there's this argument that *advanced* species would be beyond this kind of thing. But man, it only takes one. I can't believe that not even one in a billion alien civilizations would have some instinctual preference for galactic conquest for galactic conquest's own sake. I mean,

even if most humans upload themselves, there will be a couple who don't and who want to go exploring. You're trying to tell me this model applies to 999,999,999 out of one billion civilizations, and then the very first civilization we test it on, it fails?

**The Great Filter is not alien exterminators**. It sort of makes sense, from a human point of view. Maybe the first alien species to attain superintelligence was jealous, or just plain jerks, and decided to kill other species before they got the chance to catch up. Knowledgeable people like as Carl Sagan and Stephen Hawking have condemned our reverse-SETI practice of sending messages into space to see who's out there, because everyone out there may be terrible. On this view, the dominant alien civilization is the Great Filter, killing off everyone else while not leaving a visible footprint themselves.

Although I get the precautionary principle, Sagan et al's warnings against sending messages seem kind of silly to me. This isn't a failure to recognize how strong the Great Filter has to be, this is a failure to recognize how powerful a civilization that gets through it can become.

It doesn't matter one way or the other if we broadcast we're here. If there are alien superintelligences out there, *they know*. "Oh, my billion-year-old universe-spanning superintelligence wants to destroy fledgling civilizations, but we just can't find them! If only they would send very powerful radio broadcasts into space so we could figure out where they are!" No. Just no. If there are alien superintelligences out there, they tagged Earth as potential troublemakers sometime in the Cambrian Era and have been watching us very closely ever since. They know what you had for breakfast this morning and they know what Jesus had for breakfast the morning of the Crucifixion. People worried about accidentally "revealing themselves" to an intergalactic supercivilization are like Sentinel Islanders reluctant to send a message in a bottle lest modern civilization discover their existence – unaware that modern civilization has spy satellites orbiting the planet that can pick out whether or not they shaved that morning.

What about alien exterminators who are okay with weak civilizations, but kill them when they show the first sign of becoming a threat (like inventing fusion power or leaving their home solar system)? Again, you are underestimating billion-year-old universe-spanning superintelligences. Don't flatter yourself here. *You cannot threaten them*.

What about alien exterminators who are okay with weak civilizations, but destroy strong civilizations not because they feel threatened, but just for aesthetic reasons? I can't be certain that's false, but it seems to me that if they have let us continue existing this long, even though we are made of matter that can be used for something else, that has to be a conscious decision made out of something like morality. And because they're omnipotent, they have the ability to satisfy all of their (not logically contradictory) goals at once without worrying about tradeoffs. That makes me think that whatever moral impulse has driven them to allow us to survive will *probably* continue to allow us to survive even if we start annoying them for some reason. When you're omnipotent, the option of stopping the annoyance without harming anyone is just as easy as stopping the annoyance by making everyone involved suddenly vanish.

Three of these four options – x-risk, Unfriendly AI, and alien exterminators – are very very bad for humanity. I think worry about this badness has been a lot of what's driven interest in the Great Filter. I also think these are some of the least likely possible

explanations, which means we should be less afraid of the Great Filter than is generally believed.

# Book Review: Age of Em

*[Note: I really liked this book and if I criticize it that's not meant as an attack but just as what I do with interesting ideas. Note that Robin has offered to debate me about some of this and I've said no – mostly because I hate real-time debates and have bad computer hardware – but you may still want to take this into account when considering our relative positions. Mild content warning for murder, rape, and existential horror. Errors in Part III are probably my own, not the book's.]*

**I.**

There are some people who are destined to become adjectives. Pick up a David Hume book you've never read before and it's easy to recognize the ideas and style as Humean. Everything Tolkien wrote is Tolkienesque in a non-tautological sense. This isn't meant to denounce either writer as boring. Quite the opposite. They produced a range of brilliant and diverse ideas. But there was a hard-to-define and very consistent ethos at the foundation of both. Both authors were *very much like themselves*.

Robin Hanson is more like himself than anybody else I know. He's obviously brilliant – a PhD in economics, a masters in physics, work for DARPA, Lockheed, NASA, George Mason, and the Future of Humanity Institute. But his greatest aptitude is in being really, *really* Hansonian. Bryan Caplan describes it as well as anybody:

> When the typical economist tells me about his latest research, my standard reaction is 'Eh, maybe.' Then I forget about it. When Robin Hanson tells me about his latest research, my standard reaction is 'No way! Impossible!' Then I think about it for years.

This is my experience too. I think I said my first "No way! Impossible!" sometime around 2008 after reading his blog Overcoming Bias. Since then he's influenced my thinking more than almost anyone else I've ever read. When I heard he was writing a book, I was – well, I couldn't even imagine a book by Robin Hanson. When you read a thousand word blog post by Robin Hanson, you have to sit down and think about it and wait for it to digest and try not to lose too much sleep worrying about it. A whole book would be *something*.

I have now read *Age Of Em* (website)and it is indeed something. Even the cover gives you a weird sense of sublimity mixed with unease:

And in this case, judging a book by its cover is entirely appropriate.

**II.**

*Age of Em* is a work of futurism – an attempt to predict what life will be like a few generations down the road. This is not a common genre – I can't think of another book of this depth and quality in the same niche. Predicting the future is notoriously hard, and that seems to have so far discouraged potential authors and readers alike.

Hanson is not discouraged. He writes that:

> Some say that there is little point in trying to foresee the non-immediate future. But in fact there have been many successful forecasts of this sort. For example, we can reliably predict the future cost changes for devices such as batteries or solar cells, as such costs tend to follow a power law of the cumulative device

> production (Nagy et al 2013). As another example, recently a set of a thousand published technology forecasts were collected and scored for accuracy, by comparing the forecasted date of a technology milestone with its actual date. Forecasts were significantly more accurate than random, even forecasts 10 to 25 years ahead. This was true separately for forecasts made via many different methods. On average, these milestones tended to be passed a few years before their forecasted date, and sometimes forecasters were unaware that they had already passed (Charbonneau et al, 2013).
>
> A particularly accurate book in predicting the future was *The Year 2000*, a 1967 book by Herman Kahn and Anthony Wiener. It accurately predicted population, was 80% correct for computer and communication technology, and 50% correct for other technology (Albright 2002). On even longer time scales, in 1900 the engineer John Watkins did a good job of forecasting many basic features of society a century later (Watkins 1900) […]
>
> Some say no one could have anticipated the recent big changes associated with the arrival and consequences of the World Wide Web. Yet participants in the Xanadu hypertext project in which I was involved from 1984 to 1993 correctly anticipated many key aspects of the Web […] Such examples show that one can use basic theory to anticipate key elements of distant future environments, both physical and social, but also that forecasters do not tend to be much rewarded for such efforts, either culturally or materially. This helps to explain why there are relatively few serious forecasting efforst. But make no mistake, it *is* possible to forecast the future.

I think Hanson is overstating his case. All except Watkins were predicting only 10 – 30 years in the future, and most of their predictions were simple numerical estimates, eg "the population will be one billion" rather than complex pictures of society. The only project here even remotely comparable in scope to Hanson's is John Watkins' 1900 article.

Watkins is classically given some credit for broadly correct ideas like "Cameras that can send pictures across the world instantly" and "telephones that can call anywhere in the world", but of his 28 predictions, I judge only eight as even somewhat correct. For example, I grant him a prediction that "the average American will be two inches taller because of good medical care" even though he then goes on to say in the same sentence that the average life expectancy will be fifty and suburbanization will be so total that building city blocks will be illegal (sorry, John, only in San Francisco). Most of the predictions seem simply and completely false. Watkins believes all animals and insects will have been eradicated. He believes there will be "peas as large as beets" and "strawberries as large as apples" (these are two separate predictions; he is weirdly obsessed with fruit and vegetable size). We will travel to England via giant combination submarine/hovercrafts that will complete the trip in a lightning-fast two days. There will be no surface-level transportation in cities as all cars and walkways have moved underground. The letters C, X, and Q will be removed from the language. Pneumatic tubes will deliver purchases from stores. "A man or woman unable to walk ten miles at a stretch will be regarded as a weakling."

Where Watkins is right, he is generally listing a cool technology slightly beyond what was available to his time and predicting we will have it. Nevertheless, he is still mostly wrong. Yet this is Hanson's example of accurate futurology. And he is *right* to make it his example of accurate futurology, because everything else is even worse.

Hanson has no illusions of certainty. He starts by saying that "conditional on my key assumptions, I expect at least 30% of future situations to be usefully informed by my analysis. Unconditionally, I expect at least 10%." So he is not explicitly overconfident. But in an implicit sense, it's just *weird* to see the level of detail he tries to predict – for example, he has two pages about what sort of swear words the far future might use. And the book's style serves to reinforce its weirdness. The whole thing is written in a sort of professorial monotone that changes little from loving descriptions of the sorts of pipes that will cool future buildings (one of Hanson's pet topics) to speculation on our descendents' romantic relationships (key quote: "The per minute subjective value of an equal relation should not fall much below half of the per-minute value of a relation with the best available open source lover"). And it leans heavily on a favorite Hansonian literary device – the weirdly general statement about something that sounds like it can't possibly be measurable, followed by a curt reference which if followed up absolutely confirms said statement, followed by relentlessly ringing every corollary of it:

> Today, mental fatigue reduces mental performance by about 0.1% per minute. As by resting we can recover at a rate of 1% per minute, we need roughly one-tenth of our workday to be break time, with the duration between breaks being not much more than an hour or two (Trougakos and Hideg 2009; Alvanchi et al 2012)…Thus many em tasks will be designed to take about an hour, and many spurs are likely to last for about this duration.

Or:

> Today, painters, novelists, and directors who are experimental artists tend to do their best work at roughly ages 46-52, 38-50, and 45-63 respectively, but those ages are 24-34, 29-40, and 27-43, respectively for conceptual artists (Galenson 2006)…At any one time, the vast majority of actual working ems [should be] near a peak productivity subjective age.

Or:

> Wars today, like cities, are distributed evenly across all possible war sizes (Cederman 2003).

At some point I started to wonder whether Hanson was putting me on. Everything is just played *too straight*. Hanson even addresses this:

> To resist the temptation to construe the future too abstractly, I'll try to imagine a future full of complex detail. One indiciation that I've been successful in all these efforts will be if my scenario description sounds less like it came from a typical comic book or science fiction movie, and more like it came form a typical history text or business casebook.

Well, count that project a success. The effect is strange to behold, and I'm not sure it will usher in a new era of futurology. But *Age of Em* is great not *just* as futurology, but as a bunch of different ideas and purposes all bound up in a futurological package. For example:

**–** An introduction to some of the concepts that recur again and again across Robin's thought – for example, near vs. far mode, the farmer/forager dichotomy, the inside and outside views, signaling. Most of us learned these through years reading Hanson's blog *Overcoming Bias*, getting each chunk in turn, spending days or months thinking over each piece. Getting it all out of a book you can read in a couple of days sounds

really hard – but by applying them to dozens of different subproblems involved in future predictions, Hanson makes the reader more comfortable with them, and I expect a lot of people will come out of the book with an intuitive understanding of how they can be applied.

**–** A whirlwind tour through almost every science and a pretty good way to learn about the *present*. If you didn't already know that wars are distributed evenly across all possible war sizes, well, read *Age of Em* and you will know that and many similar things besides.

**–** A manifesto. Hanson often makes predictions by assuming that since the future will be more competitive, future people are likely to converge toward optimal institutions. This is a dangerous assumption for futurology – it's the same line of thinking that led Watkins to assume English would abandon C, X, and Q as inefficient – but it's a *great* assumption if you want a chance to explain your ideas of optimal institutions to thousands of people who think they're reading fun science-fiction. Thus, Robin spends several pages talking about how ems may use prediction markets – an information aggregation technique he invented – to make their decisions. In the real world, Hanson has been trying to push these for decades, with varying levels of success. Here, in the guise of a future society, he can expose a whole new group of people to their advantages – as well as the advantages of something called "combinatorial auctions" which I am still not smart enough to understand.

**–** A mind-expanding drug. One of the great risks of futurology is to fail to realize how *different* societies and institutions can be – the same way uncreative costume designers make their aliens look like humans with green skin. A lot of our thoughts about the future involve assumptions we've never really examined critically, and Hanson dynamites those assumptions. For page after page, he gives strong arguments why our descendants might be poorer, shorter-lived, less likely to travel long distances or into space, less progressive and open-minded. He predicts little noticeable technological change, millimeter-high beings living in cities the size of bottles, careers lasting fractions of seconds, humans being incomprehensibly wealthy patrons to their own robot overlords. And *all of it makes sense*.

When I read Stross' *Accelerando*, one of the parts that stuck with me the longest were the Vile Offspring, weird posthuman entities that operated a mostly-incomprehensible Economy 2.0 that humans just sort of hung out on the edges of, goggle-eyed. It was a weird vision – but, for Stross, mostly a black box. *Age of Em* opens the box and shows you every part of what our weird incomprehensible posthuman descendents will be doing in loving detail. Even what kind of swear words they'll use.

**III.**

So, what is the Age of Em?

According to Hanson, AI is really hard and won't be invented in time to shape the posthuman future. But sometime a century or so from now, scanning technology, neuroscience, and computer hardware will advance enough to allow emulated humans, or "ems". Take somebody's brain, scan it on a microscopic level, and use this information to simulate it neuron-by-neuron on a computer. A good enough simulation will map inputs to outputs in exactly the same way as the brain itself, effectively uploading the person to a computer. Uploaded humans will be much the same as biological humans. Given suitable sense-organs, effectuators, virtual avatars, or even

robot bodies, they can think, talk, work, play, love, and build in much the same way as their "parent". But ems have three very important differences from biological humans.

First, they have no natural body. They will never need food or water; they will never get sick or die. They can live entirely in virtual worlds in which any luxuries they want – luxurious penthouses, gluttonous feasts, Ferraris – can be conjured out of nothing. They will have some limited ability to transcend space, talking to other ems' virtual presences in much the same way two people in different countries can talk on the Internet.

Second, they can run at different speeds. While a normal human brain is stuck running at the speed that physics allow, a computer simulating a brain can simulate it faster or slower depending on preference and hardware availability. With enough parallel hardware, an em could experience a subjective century in an objective week. Alternatively, if an em wanted to save hardware it could process all its mental operations *v e r y s l o w l y* and experience only a subjective week every objective century.

Third, just like other computer data, ems can be copied, cut, and pasted. One uploaded copy of Robin Hanson, plus enough free hardware, can become a thousand uploaded copies of Robin Hanson, each living in their own virtual world and doing different things. The copies could even converse with each other, check each other's work, duel to the death, or – yes – have sex with each other. And if having a thousand Robin Hansons proves too much, a quick ctrl-x and you can delete any redundant ems to free up hard disk space for Civilization 6 (coming out this October!)

Would this count as murder? Hanson predicts that ems will have unusually blase attitudes toward copy-deletion. If there are a thousand other copies of me in the world, then going to sleep and not waking up just feels like delegating back to a different version of me. If you're still not convinced, Hanson's essay Is Forgotten Party Death? is a typically disquieting analysis of this proposition. But whether it's true or not is almost irrelevant – at least *some* ems will think this way, and they will be the ones who tend to volunteer to be copied for short term tasks that require termination of the copy afterwards. If you personally aren't interested in participating, the economy will leave you behind.

The ability to copy ems as many times as needed fundamentally changes the economy and the idea of economic growth. Imagine Google has a thousand positions for Ruby programmers. Instead of finding a thousand workers, they can find one very smart and very hard-working person and copy her a thousand times. With unlimited available labor supply, wages plummet to subsistence levels. "Subsistence levels" for ems are the bare minimum it takes to rent enough hardware from Amazon Cloud to run an em. The overwhelming majority of ems will exist at such subsistence levels. On the one hand, if you've got to exist on a subsistence level, a virtual world where all luxuries can be conjured from thin air is a pretty good place to do it. On the other, such starvation wages might leave ems with little or no leisure time.

Sort of. This gets weird. There's an urban legend about a "test for psychopaths". You tell someone a story about a man who attends his mother's funeral. He met a really pretty girl there and fell in love, but neglected to get her contact details before she disappeared. How might he meet her again? If they answer "kill his father, she'll probably come to that funeral too", they're a psychopath – ordinary people would have a mental block that prevents them from even considering such a drastic solution. And I bring this up because after reading *Age of Em* I feel like Robin Hanson would be

able to come up with some super-solution even the psychopaths can't think of, some plan that gets the man a threesome with the girl and her even hotter twin sister at the cost of wiping out an entire continent. Everything about labor relations in *Age of Em* is like this.

For example, suppose you want to hire an em at subsistence wages, but you want them 24 hours a day, 7 days a week. Ems probably need to sleep – that's hard-coded into the brain, and the brain is being simulated at enough fidelity to leave that in. But jobs with tasks that don't last longer than a single day – for example, a surgeon who performs five surgeries a day but has no day-to-day carryover – can get around this restriction by letting an em have one full night of sleep, then copying it. Paste the em at the beginning of the workday. When it starts to get tired, let it finish the surgery it's working on, then delete it and paste the well-rested copy again to do the next surgery. Repeat forever and the em never has to get any more sleep than that one night. You can use the same trick to give an em a "vacation" – just give it *one* of them, then copy-paste that brain-state forever.

Or suppose your ems want frequent vacations, but you want them working every day. Let a "trunk" em vacation every day, then make a thousand copies every morning, work all the copies for twenty-four hours, then delete them. Every copy remembers a life spent in constant vacation, and cheered on by its generally wonderful existence it will give a full day's work. But from the company's perspective, 99.9% of the ems in its employment are working at any given moment.

(another option: work the em at normal subjective speed, then speed it up a thousand times to take its week-long vacation, then have it return to work after only one-one-thousandth of a week has passed in real life)

Given that ems exist at subsistence wages, saving enough for retirement sounds difficult, but this too has weird psychopathic solutions. Thousands of copies of the same em can pool their retirement savings, then have all except a randomly chosen one disappear at the moment of retirement, leaving that one with an nest egg thousands of time what it could have accumulated by its own efforts. Or an em can invest its paltry savings in some kind of low-risk low-return investment and reduce its running speed so much that the return on its investment is enough to pay for its decreased subsistence. For example, if it costs $100 to rent enough computing power to run an em at normal speed for one year, and you only have $10 in savings, you can rent 1/1000th of the computer for $0.10, run at 1/1000th speed, invest your $10 in a bond that pays 1% per year, and have enough to continue running indefinitely. The only disadvantage is that you'll only experience a subjective week every twenty objective years. Also, since other entities are experiencing a subjective week every second, and some of those entities have nukes, probably there will be some kind of big war, someone will nuke Amazon's data centers, and you'll die after a couple of your subjective minutes. But at least you got to retire!

If ems do find ways to get time off the clock, what will they do with it? Probably they'll have really weird social lives. After all, the existence of em copies is mostly funded by companies, and there's no reason for companies to copy-paste any but the best workers in a given field. So despite the literally trillions of ems likely to make up the world, most will be copies of a few exceptionally brilliant and hard-working individuals with specific marketable talents. Elon Musk might go out one day to the bar with his friend, who is also Elon Musk, and order "the usual". The bartender, who is Elon Musk himself, would know exactly what drink he wants and have it readily available, as the bar caters entirely to people who are Elon Musk. A few minutes later, a few Chesley

Sullenbergers might come in after a long day of piloting airplanes. Each Sullenberger would have met hundreds of Musks before and have a good idea about which Musk-Sullenberger conversation topics were most enjoyable, but they might have to adjust for circumstances; maybe the Musks they met before all branched off a most recent common ancestor in 2120, but these are a different branch who were created in 2105 and remember Elon's human experiences but not a lot of the posthuman lives that shaped the 2120 Musks' worldviews. One Sullenberger might tentatively complain that the solar power grid has too many outages these days; a Musk might agree to take the problem up with the Council of Musks, which is totally a thing that exist (Hanson calls these sorts of groups "copy clans" and says they are "a natural candidate unit for finance, reproduction, legal, liability, and political representation").

Romance could be even weirder. Elon Musk #2633590 goes into a bar and meets Taylor Swift #105051, who has a job singing in a nice local nightclub and so is considered prestigious for a Taylor Swift. He looks up a record of what happens when Elon Musks ask Taylor Swifts out and finds they are receptive on 87.35% of occasions. The two start dating and are advised by the Council of Musks and the Council of Swifts on the issues that are known to come up in Musk-Swift relationships and the best solutions that have been found to each. Unfortunately, Musk #2633590 is transferred to a job that requires operating at 10,000x human speed, but Swift #105051's nightclub runs at 100x speed and refuses to subsidize her to run any faster; such a speed difference makes normal interaction impossible. The story has a happy ending; Swift #105051 allows Musk #2633590 to have her source code, and whenever he is feeling lonely he spends a little extra money to instantiate a high-speed copy of her to hang out with.

(needless to say, these examples are not exactly word-for-word taken from the book, but they're heavily based off of Hanson's more abstract descriptions)

The em world is not just very weird, it's also very very big. Hanson notes that labor is a limiting factor in economic growth, yet even today the economy doubles about once every fifteen years. Once you can produce skilled labor through a simple copy-paste operation, especially labor you can run at a thousand times human speed, the economy will go through the roof. He writes that:

> To generate an empirical estimate of em economy doubling times, we can look at the timescales it takes for machine shopes and factories today to make a mass of machines of a quality, quantity, variety, and value similar to that of machines that they themselves contain. Today that timescale is roughly 1 to 3 months. Also, designs were sketched two to three decades ago for systems that might self-repliate nearly completeld in 6 to 12 months…these estimates suggest that today's manufacturing technologiy is capable of self-repliating on a scale of a few weeks to a few months.

Hanson thinks that with further innovation, such times can be reduced so far that "the economy might double every objective year, month, week, or day." As the economy doubles the labor force – ie the number of ems – may double with it, until only a few years after the first ems the population numbers in the trillions. But if the em population is doubling every day, there had better be some pretty amazing construction efforts going on. The only thing that could possibly work on that scale is prefabricated modular construction of giant superdense cities, probably made mostly out of some sort of proto early-stage computronium (plus cooling pipes). Ems would be reluctant to travel from one such city to another – if they exist at a thousand times human speed, a trip on a hypersonic airliner that could go from New York to Los

Angeles in an hour would still take forty subjective days. Who wants to be on an airplane for forty days?

(long-distance trade is also rare, since if the economy doubles fast enough it means that by the time goods reach their destination they could be almost worthless)

The real winners of this ultra-fast-growing economy? Ordinary humans. While humans will be way too slow and stupid to do anything useful, they will tend to have non-subsistence amounts of money saved up from their previous human lives, and also be running at speeds thousands of times slower than most of the economy. When the economy doubles every day, so can your bank account. Ordinary humans will become rarer, less relevant, but fantastically rich – a sort of doddering Neanderthal aristocracy spending sums on a cheeseburger that could support thousands of ems in luxury for entire lifetimes. While there will no doubt be pressure to liquidate humans and take their stuff, Hanson hopes that the spirit of rule of law – the same spirit that protects rich minority groups today – will win out, with rich ems reluctant to support property confiscation lest it extend to them also. Also, em retirees will have incentives a lot like humans – they have saved up money and go really slow – and like AARP memembers today they may be able to obtain disproportionate political power which will then protect the interests of slow rich people.

But we might not have much time to enjoy our sudden rise in wealth. Hanson predicts that the Age of Em will last for subjective em millennia – ie about one to two actual human years. After all, most of the interesting political and economic activity is going on at em timescales. In the space of a few subjective millennia, either someone will screw up and cause the apocalypse, somebody will invent real superintelligent AI that causes a technological singularity, or some other weird thing will happen taking civilization beyond the point that even Robin dares to try to predict.

**IV.**

Hanson understands that people might not like the idea of a future full of people working very long hours at subsistence wages forever (Zack Davis' Contract-Drafting Em song is, as usual, relevant). But Hanson himself does not view this future as dystopian. Despite our descendents' by-the-numbers poverty, they will avoid the miseries commonly associated with poverty today. There will be no dirt or cockroaches in their sparkling virtual worlds, nobody will go hungry, petty crime will be all-but-eliminated, and unemployment will be low. Anybody who can score some leisure time will have a dizzying variety of hyperadvanced entertainment available, and as for the people who can't, they'll mostly have been copied from people who really like working hard and don't miss it anyway. As unhappy as we moderns may be contemplating em society, ems themselves will not be unhappy! And as for us:

> The analysis in this book suggests that lives in the next great era may be as different from our lives as our lives are from farmers' lives, or farmers' lives are from foragers' lives. Many readers of this book, living industrial era lives and sharing industrial era values, may be disturbed to see a forecast of em era descendants with choices and lifestyles that appear to reject many of the values that they hold dear. Such readers may be tempted to fight to prevent the em future, perhaps preferring a continuation of the industrial era. Such readers may be correct that rejecting the em future holds them true to their core values. But I advise such readers to first try hard to see this new era in some detail from the point of view of its typical residents. See what they enjoy and what fills them with pride, and listen to their criticisms of your era and values.

A short digression: there's a certain strain of thought I find infuriating, which is "My traditionalist ancestors would have disapproved of the changes typical of my era, like racial equality, more open sexuality, and secularism. But I am smarter than them, and so totally okay with how the future will likely have values even more progressive and shocking than my own. Therefore I pre-approve of any value changes that might happen in the future as definitely good and better than our stupid hidebound present."

I once read a science-fiction story that depicted a pretty average sci-fi future – mighty starships, weird aliens, confederations of planets, post-scarcity economy – with the sole unusual feature that rape was considered totally legal, and opposition to such as bigoted and ignorant as opposition to homosexuality is today. Everybody got really angry at the author and said it was offensive for him to even speculate about that. Well, that's the method by which our cheerful acceptance of any possible future values is maintained: restricting the set of "any possible future values" to "values slightly more progressive than ours" and then angrily shouting down anyone who discusses future values that actually sound bad. But of course the *whole question* of how worried to be about future value drift *only makes sense* in the context of future values that genuinely violate our current values. Approving of all future values except ones that would be offensive to even speculate about is the same faux-open-mindedness as tolerating anything except the outgroup.

Hanson deserves credit for positing a future whose values are likely to upset even the sort of people who say they don't get upset over future value drift. I'm not sure whether or not he deserves credit for not being upset by it. Yes, it's got low-crime, ample food for everybody, and full employment. But so does *Brave New World*. The whole *point* of dystopian fiction is pointing out that we have complicated values beyond material security. Hanson is absolutely right that our traditionalist ancestors would view our own era with as much horror as some of us would view an em era. He's even right that on utilitarian grounds, it's hard to argue with an em era where everyone is really happy working eighteen hours a day for their entire lives because we selected for people who feel that way. But at some point, can we make the Lovecraftian argument of "I know my values are provincial and arbitrary, but they're *my* provincial arbitrary values and I will make any sacrifice of blood or tears necessary to defend them, even unto the gates of Hell?"

This brings us to an even worse scenario.

There are a lot of similarities between Hanson's futurology and (my possibly erroneous interpretation of) the futurology of Nick Land. I see Land as saying, like Hanson, that the future will be one of quickly accelerating economic activity that comes to dominate a bigger and bigger portion of our descendents' lives. But whereas Hanson's framing focuses on the participants in such economic activity, playing up their resemblances with modern humans, Land takes a bigger picture. He talks about the economy itself acquiring a sort of self-awareness or agency, so that the destiny of civilization is consumed by the imperative of economic growth.

Imagine a company that manufactures batteries for electric cars. The inventor of the batteries might be a scientist who really believes in the power of technology to improve the human race. The workers who help build the batteries might just be trying to earn money to support their families. The CEO might be running the business because he wants to buy a really big yacht. And the whole thing is there to eventually, somewhere down the line, let a suburban mom buy a car to take her kid to soccer practice. Like most companies the battery-making company is primarily a profit-

making operation, but the profit-making-ness draws on a lot of not-purely-economic actors and their not-purely-economic subgoals.

Now imagine the company fires all its employees and replaces them with robots. It fires the inventor and replaces him with a genetic algorithm that optimizes battery design. It fires the CEO and replaces him with a superintelligent business-running algorithm. All of these are good decisions, from a profitability perspective. We can absolutely imagine a profit-driven shareholder-value-maximizing company doing all these things. But it reduces the company's non-masturbatory participation in an economy that points outside itself, limits it to just a tenuous connection with soccer moms and maybe some shareholders who want yachts of their own.

Now take it further. Imagine there are no human shareholders who want yachts, just banks who lend the company money in order to increase their own value. And imagine there are no soccer moms anymore; the company makes batteries for the trucks that ship raw materials from place to place. Every non-economic goal has been stripped away from the company; it's just an appendage of Global Development.

Now take it even further, and imagine this is what's happened everywhere. There are no humans left; it isn't economically efficient to continue having humans. Algorithm-run banks lend money to algorithm-run companies that produce goods for other algorithm-run companies and so on ad infinitum. Such a masturbatory economy would have all the signs of economic growth we have today. It could build itself new mines to create raw materials, construct new roads and railways to transport them, build huge factories to manufacture them into robots, then sell the robots to whatever companies need more robot workers. It might even eventually invent space travel to reach new worlds full of raw materials. Maybe it would develop powerful militaries to conquer alien worlds and steal their technological secrets that could increase efficiency. It would be vast, incredibly efficient, and utterly pointless. The real-life incarnation of those strategy games where you mine Resources to build new Weapons to conquer new Territories from which you mine more Resources and so on forever.

But this seems to me the natural end of the economic system. Right now it needs humans only as laborers, investors, and consumers. But robot laborers are potentially more efficient, companies based around algorithmic trading are already pushing out human investors, and most consumers already aren't individuals – they're companies and governments and organizations. At each step you can gain efficiency by eliminating humans, until finally humans aren't involved *anywhere*.

True to form, Land doesn't see this as a dystopia – I think he conflates "maximally efficient economy" with "God", which is a *hell* of a thing to conflate – but I do. And I think it provides an important new lens with which to look at the Age of Em.

The Age of Em is an economy in the early stages of such a transformation. Instead of being able to replace everything with literal robots, it replaces them with humans who have had some aspects of their humanity stripped away. Biological bodies. The desire and ability to have children normally. Robin doesn't think people will lose all leisure time and non-work-related desires, but he doesn't seem too sure about this and it doesn't seem to bother him much if they do.

I envision a spectrum between the current world of humans and Nick Land's Ascended Economy. Somewhere on the spectrum we have ems who get leisure time. A little further on the spectrum we have ems who don't get leisure time.

But we can go further. Hanson imagines that we can "tweak" em minds. We may not understand the brain enough to create totally new intelligences from the ground up, but by his Age of Em we should understand it well enough to make a few minor hacks, the same way even somebody who doesn't know HTML or CSS can usually figure out how to change the background color of a webpage with enough prodding. Many of these mind tweaks will be the equivalent of psychiatric drugs – some might even be computer simulations of what we observe to happen when we give psychiatric drugs to a biological brain. But these tweaks will necessarily be much stronger and more versatile, since we no longer care about bodily side effects (ems don't have bodies) and we can apply it to only a single small region of the brain and avoid actions anywhere else. You could also very quickly advance brain science – the main limits today are practical (it's really hard to open up somebody's brain and do stuff to it without killing them) and ethical (the government might have some words with you if you tried). An Age of Em would remove both obstacles, and give you the added bonus of being able to make thousands of copies of your test subjects for randomized controlled trials, reloading any from a saved copy if they died. Hanson envisions that:

> As the em world is a very competitive world where sex is not needed for reproduction, and as sex can be time and attention-consuming, ems may try to suppress sexuality, via mind tweaks that produce effects analogous to castration. Such effects might be temporary, perhaps with a consciously controllable on-off switch…it is possible that em brain tweaks could be found to greatly reduce natural human desires for sex and related romantic and intimate pair bonding without reducing em productivity. It is also possible that many of the most productive ems would accept such tweaks.

Possible? I can do that *right now* with a high enough dose of Paxil, and I don't even have to upload your brain to a computer first. Fun stories about Musk #2633590 and Swift #105051 aside, I expect this would happen about ten minutes after the advent of the Age of Em, and we would have taken another step down the path to the Ascended Economy.

There are dozens of other such tweaks I can think of, but let me focus on two.

First, stimulants have a very powerful ability to focus the brain on the task at hand, as anybody who's taken Adderall or modafinil can attest. Their main drawbacks are addictiveness and health concerns, but in a world where such pills can be applied as mental tweaks, where minds have no bodies, and where any mind that gets too screwed up can be reloaded from a backup copy, these are barely concerns at all. Many of the purely mental side effects of stimulants come from their effects in parts of the brain not vital to the stimulant effect. If we can selectively apply Adderall to certain brain centers but not others, then unapply it at will, then from employers' point of view there's no reason not to have all workers dosed with superior year 2100 versions of Adderall at all times. I worry that not only will workers not have any leisure time, but they'll be neurologically incapable of having their minds drift off while on the job. Davis' contract-drafting em who starts wondering about philosophy on the job wouldn't get terminated. He would just have his simulated-Adderall dose increased.

Second, Robin managed to write an entire book about emulated minds without using the word "wireheading". This is another thing we can do right now, with today's technology – but once it's a line of code and not a costly brain surgery, it should become nigh-universal. Give ems the control switches to their own reward centers and all questions about leisure time become irrelevant. Give *bosses* the control switches to their employees' reward centers, and the situation changes markedly. Hanson says

that there probably won't be too much slavery in the em world, because it will likely have strong rule of law, because slaves aren't as productive as free workers, and there's little advantage to enslaving someone when you could just pay them subsistence wages anyway. But slavery isn't *nearly* as abject and inferior a condition as the one where somebody else has the control switch to your reward center. Combine that with the stimulant use mentioned above, and you can have people who will never have nor want to have any thought about anything other than working on the precise task at which they are supposed to be working at any given time.

This is something I worry about even in the context of normal biological humans. But Hanson already believes em worlds will have few regulations and be able to ignore the moral horror of 99% of the population by copying and using the 1% who are okay with something. Combine this with a situation where brains are easily accessible and tweakable, and this sort of scenario becomes horribly likely.

I see almost no interesting difference between an em world with full use of these tweaks and an Ascended Economy world. Yes, there are things that look vaguely human in outline laboring in the one and not the other, but it's not like there will be different thought processes or different results. I'm not even sure what it would mean for the ems to be conscious in a world like this – they're not doing anything interesting with the consciousness. The best we could say about this is that if the wireheading is used liberally it's a lite version of the world where everything gets converted to hedonium.

## V.

In a book full of weird ideas, there is only one idea rejected as too weird. And in a book written in a professorial monotone, there's only one point at which Hanson expresses anything like emotion:

> Some people foresee a rapid local "intelligence explosion" happening soon after a smart AI system can usefully modify its local architecture (Chalmers 2010; Hanson and Yudkowsky 2013; Yudkowsky 2013; Bostrom 2014)…Honestly to me this local intelligence explosion scenario looks suspiciously like a super-villain comic book plot. A flash of insight by a lone genius lets him create a genius AI. Hidden in its super-villain research lab lair, this guines villain AI works out unprecedented revolutions in AI design, turns itself into a super-genius, which then invents super-weapons and takes over the world. Bwa ha ha.

For someone who just got done talking about the sex lives of uploaded computers in millimeter-tall robot bodies running at 1000x human speed, Robin is sure quick to use the absurdity heuristic to straw-man intelligence explosion scenarios as "comic book plots". Take away his weird authorial tic of using the words "genius" and "supervillain", this scenario reduces to "Some group, perhaps Google, perhaps a university, invent an artificial intelligence smart enough to edit its own source code; exponentially growing intelligence without obvious bound follows shortly thereafter". Yes, it's weird to think that there may be a sudden quantum leap in intelligence like this, but no weirder than to think most of civilization will transition from human to em in the space of a year or two. I'm a little bit offended that this is the only idea given this level of dismissive treatment. Since I do have immense respect for Robin, I hope my offense doesn't color the following thoughts too much.

Hanson's arguments against AI seem somewhat motivated. He admits that AI researchers generally estimate less than 50 years before we get human-level artificial

intelligence, a span shorter than his estimate of a century until we can upload ems. He even admits that no AI researcher thinks ems are a plausible route to AI. But he dismisses this by saying when he asks AI experts informally, they say that in their own field, they have only noticed about 5-10% of the progress they expect would be needed to reach human intelligence over the past twenty years. He then multiplies out to say that it will probably take at least 400 years to reach human-level AI. I have two complaints about this estimate.

First, he is explicitly ignoring published papers surveying hundreds of researchers using validated techniques, in favor of what he describes as "meeting experienced AI experts informally". But even though he feels comfortable rejecting vast surveys of AI experts as potentially biased, as best I can tell he does not ask a single neuroscientist to estimate the date at which brain scanning and simulation might be available. He just says that "it seems plausible that sufficient progress will be made in roughly a century or so", citing a few hopeful articles by very enthusiastic futurists who are not neuroscientists or scanning professionals themselves and have not talked to any. This seems to me to be an extreme example of isolated demands for rigor. No matter how many AI scientists think AI is soon, Hanson will cherry-pick the surveying procedures and results that make it look far. But if a few futurists think brain emulation is possible, then no matter what anybody else thinks that's good enough for him.

Second, one would expect that even if there were only 5-10% progress over the last twenty years, then there would be faster progress in the future, since the future will have a bigger economy, better supporting technology, and more resources invested in AI research. Robin answers this objection by saying that "increases in research funding usually give much less than proportionate increases in research progress" and cites Alston et al 2011. I looked up Alston et al 2011, and it is a paper relating crop productivity to government funding of agriculture research. There was no attempt to relate its findings to any field other than agriculture, nor to any type of funding other than government. But studies show that while public research funding often does have minimal effects, the effect of private research funding is usually much larger. A single sentence citing a study in crop productivity to apply to artificial intelligence while ignoring much more relevant results that contradict it seems like a really weak argument for a statement as potentially surprising as "amount of research does not affect technological progress".

I realize that Hanson has done a lot more work on this topic and he couldn't fit all of it in this book. I disagree with his other work too, and I've said so elsewhere. For now I just want to say that the arguments in this book seem weak to me.

I also want to mention what seems to me a very Hansonian counterargument to the ems-come-first scenario: we have always developed de novo technology before understanding the relevant biology. We built automobiles by figuring out the physics of combustion engines, not by studying human muscles and creating mechanical imitations of myosin and actin. Although the Wright brothers were inspired by birds, their first plane was not an ornithopter. Our power plants use coal and uranium instead of the Krebs Cycle. Biology is *really hard*. Even slavishly *copying* biology is really hard. I don't think Hanson and the futurists he cites understand the scale of the problem they've set themselves.

Current cutting-edge brain emulation projects have found their work much harder than expected. Simulating a nematode is pretty much the rock-bottom easiest thing in this category, since they are tiny primitive worms with only a few neurons; the history of the field is a litany of failures, with current leader OpenWorm "reluctant to make

bold claims about its current resemblance to biological behavior". A more ambitious $1.3 billion attempt to simulate a tiny portion of a rat brain has gone down in history as a legendary failure (politics were involved, but I expect they would be involved in a plan to upload a human too). And these are just attempts to get something that behaves *vaguely* like a nematode or rat. Actually uploading a human, keeping their memory and personality intact, and not having them go insane afterwards boggles the mind. We're still not sure how much small molecules matter to brain function, how much glial cells matter to brain function, how many things in the brain are or aren't local. AI researchers are making programs that can defeat chess grandmasters; upload researchers are still struggling to make a worm that will wriggle. The right analogy for modern attempts to upload human brains isn't modern attempts at designing AI. It's an attempt at designing AI by someone who doesn't even know how to plug in a computer.

**VI.**

I guess what really bothers me about Hanson's pooh-poohing of AI is him calling it "a comic book plot". To me, it's Hanson's scenario that seems science-fiction-ish.

I say this not as a generic insult but as a pointer at a specific category of errors. In *Star Wars*, the Rebellion had all of these beautiful hyperspace-capable starfighters that could shoot laser beams and explore galaxies – and *they still had human pilots*. 1977 thought the pangalactic future would still be using people to pilot its military aircraft; in reality, even 2016 is moving away from this.

Science fiction books have to tell interesting stories, and interesting stories are about humans or human-like entities. We can enjoy stories about aliens or robots as long as those aliens and robots are still approximately human-sized, human-shaped, human-intelligence, and doing human-type things. A Star Wars in which all of the X-Wings were combat drones wouldn't have done anything for us. So when I accuse something of being science-fiction-ish, I mean bending over backwards – and ignoring the evidence – in order to give basically human-shaped beings a central role.

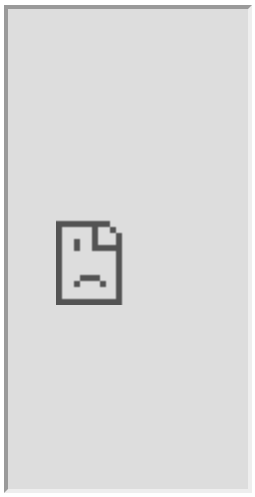
This is my critique of Robin. As weird as the Age of Em is, it makes sure never to be weird in ways that warp the fundamental humanity of its participants. Ems might be copied and pasted like so many .JPGs, but they still fall in love, form clans, and go on vacations.

In contrast, I expect that we'll get some kind of AI that will be totally inhuman and much harder to write sympathetic stories about. If we get ems after all, I expect them to be lobotomized and drugged until they become *effectively* inhuman, cogs in the Ascended Economy that would no more fall in love than an automobile would eat hay and whinny. Robin's interest in keeping his protagonists relatable makes his book fascinating, engaging, and probably wrong.

I almost said "and probably less horrible than we should actually expect", but I'm not sure that's true. With a certain amount of horror-suppressing, the Ascended Economy can be written off as morally neutral – either having no conscious thought, or stably wireheaded. All of Robin's points about how normal non-uploaded humans should be able to survive an Ascended Economy at least for a while seem accurate. So morally valuable actors might continue to exist in weird Amish-style enclaves, living a post-scarcity lifestyle off the proceeds of their investments, while all the while the Ascended Economy buzzes around them, doing weird inhuman things that encroach

upon them not at all. This seems slightly worse than a Friendly AI scenario, but much better than we have any right to expect of the future.

I highly recommend *Age of Em* as a fantastically fun read and a great introduction to these concepts. It's engaging, readable, and *weird*. I just don't know if it's weird *enough*.



# Ascended Economy?

*[Obviously speculative futurism is obviously speculative. Complex futurism may be impossible and I should feel bad for doing it anyway. This is "inspired by" Nick Land – I don't want to credit him fully since I may be misinterpreting him, and I also don't want to avoid crediting him at all, so call it "inspired".]*

## I.

My review of Age of Em mentioned the idea of an "ascended economy", one where economic activity drifted further and further from human control until finally there was no relation at all. Many people rightly questioned that idea, so let me try to expand on it further. What I said there, slightly edited for clarity:

> Imagine a company that manufactures batteries for electric cars. The inventor of the batteries might be a scientist who really believes in the power of technology to improve the human race. The workers who help build the batteries might just be trying to earn money to support their families. The CEO might be running the business because he wants to buy a really big yacht. The shareholders might be holding the stock to help save for a comfortable retirement. And the whole thing is there to eventually, somewhere down the line, let a suburban mom buy a car to take her kid to soccer practice. Like most companies the battery-making company is primarily a profit-making operation, but the profit-making-ness draws on a lot of not-purely-economic actors and their not-purely-economic subgoals.
>
> Now imagine the company fires the inventor and replaces him with a genetic algorithm that optimizes battery design. It fires all its employees and replaces them with robots. It fires the CEO and replaces him with a superintelligent business-running algorithm. All of these are good decisions, from a profitability perspective. We can absolutely imagine a profit-driven shareholder-value-maximizing company doing all these things. But it reduces the company's non-masturbatory participation in an economy that points outside itself, limits it to just a tenuous connection with soccer moms and maybe some shareholders who want yachts of their own.
>
> Now take it further. Imagine that instead of being owned by humans directly, it's owned by an algorithm-controlled venture capital fund. And imagine there are no soccer moms anymore; the company makes batteries for the trucks that ship raw materials from place to place. Every non-economic goal has been stripped away from the company; it's just an appendage of Global Development.
>
> Now take it even further, and imagine this is what's happened everywhere. Algorithm-run banks lend money to algorithm-run companies that produce goods for other algorithm-run companies and so on ad infinitum. Such a masturbatory economy would have all the signs of economic growth we have today. It could build itself new mines to create raw materials, construct new roads and railways to transport them, build huge factories to manufacture them into robots, then sell the robots to whatever companies need more robot workers. It might even eventually invent space travel to reach new worlds full of raw materials. Maybe it would develop powerful militaries to conquer alien worlds and steal their technological secrets that could increase efficiency. It would be vast, incredibly efficient, and utterly pointless. The real-life incarnation of those strategy games where you mine Resources to build new Weapons to conquer new Territories from which you mine more Resources and so on forever.

This is obviously weird and I probably went too far, but let me try to explain my reasoning.

The part about replacing workers with robots isn't too weird; lots of industries have already done that. There's a whole big debate over to what degree that will intensify, and whether unemployed humans will find jobs somewhere else, or whether there will only be jobs for creative people with a certain education level or IQ. This part is well-discussed and I don't have much to add.

But lately there's also been discussion of automating corporations themselves. I don't know much about Ethereum (and I probably shouldn't guess since I think the inventor reads this blog and could call me on it) but as I understand it they aim to replace corporate governance with algorithms. For example, the DAO is a leaderless investment fund that allocates money according to member votes. Right now this isn't *super* interesting; algorithms can't make too many difficult business decisions so it's limited to corporations that just do a couple of primitive actions (and why would anyone want a democratic venture fund?). But once we get closer to true AI, they might be able to make the sort of business decisions that a CEO does today. The end goal is intelligent corporations controlled by nobody but themselves.

This very blog has an advertisement for a group trying to make investment decisions based on machine learning. If they succeed, how long is it before some programmer combines a successful machine investor with a DAO-style investment fund, and creates an entity that takes humans out of the loop completely? You send it your money, a couple years later it gives you back hopefully more money, with no humans involved at any point. Such robo-investors might eventually become more efficient than Wall Street – after all, hedge fund managers get super rich by skimming money off the top, and any entity that doesn't do that would have an advantage above and beyond its investment acumen.

If capital investment gets automated, corporate governance gets automated, and labor gets automated, we might end up with the creepy prospect of ascended corporations – robot companies with robot workers owned by robot capitalists. Humans could become irrelevant to most economic activity. Run such an economy for a few hundred years and what do you get?

**II.**

But *in the end* isn't all this about humans? Humans as the investors giving their money to the robo-venture-capitalists, then reaping the gains of their success? And humans as the end consumers whom everyone is eventually trying to please?

It's possible to imagine accidentally forming stable economic loops that don't involve humans. Imagine a mining-robot company that took one input (steel) and produced one output (mining-robots), which it would sell either for money or for steel below a certain price. And imagine a steel-mining company that took one input (mining-robots) and produced one output (steel) which it would sell for either money or for mining-robots below a certain price. The two companies could get into a stable loop and end up tiling the universe with steel and mining-robots without caring whether anybody else wanted either. Obviously the real economy is a zillion times more complex than that, and I'm nowhere near the level of understanding I would need to say if there's any chance that an entire self-sustaining economy worth of things could produce a loop like that. But I guess you only need one.

I think we can get around this in a causal-historical perspective, where we start with only humans and no corporations. The first corporations that come into existence have to be those that want to sell goods to humans. The next level of corporations can be those that sell goods to corporations that sell to humans. And so on. So unless a stable loop forms by accident, all corporations should exist to serve humans. A sufficiently rich human could finance the creation of a stable loop if they wanted to, but why would they want to? Since corporations exist only to satisfy human demand on some level or another, and there's no demand for stable loops, corporations wouldn't finance the development of stable loops, except by accident.

(for an interesting accidental stable loop, check out this article on the time two bidding algorithms accidentally raised the price of a book on fly genetics to more than $20 million)

Likewise, I think humans should always be the stockholders of last resort. Since humans will have to invest in the first corporation, even if that corporation invests in other corporations which invest in other corporations in turn, eventually it all bottoms down in humans (is this right?)

The only way I can see humans being eliminated from the picture is, again, by accident. If there are a hundred layers between some raw material corporation and humans, then if each layer is *slightly* skew to what the layer below it wants, the hundredth layer could be really *really* skew. Theoretically all our companies *today* are grounded in serving the needs of humans, but people are still thinking of spending millions of dollars to build floating platforms exactly halfway between New York and London in order to exploit light-speed delays to arbitrage financial markets better, and I'm not sure which human's needs that serves exactly. I don't know if there are bounds to how much of an economy can be that kind of thing.

Finally, humans might deliberately create small nonhuman entities with base level "preferences". For example, a wealthy philanthropist might create an ascended charitable organization which supports mathematical research. Now 99.9% of base-level preferences guiding the economy would be human preferences, and 0.1% might be a hard-coded preference for mathematics research. But since non-human agents at the base of the economy would only be as powerful as the proportion of the money supply they hold, most of the economy would probably still overwhelmingly be geared towards humans unless something went wrong.

Since the economy could grow much faster than human populations, the economy-to-supposed-consumer ratio might become so high that things start becoming ridiculous. If the economy became a light-speed shockwave of economium (a form of matter that maximizes shareholder return, by analogy to computronium and hedonium) spreading across the galaxy, how does all that productive power end up serving the same few billion humans we have now? It would probably be really wasteful, the cosmic equivalent of those people who specialize in getting water from specific glaciers on demand for the super-rich because the super-rich can't think of anything better to do with their money. Except now the glaciers are on Pluto.

**III.**

Glacier water from Pluto sounds pretty good. And we can hope that things will get so post-scarcity that governments and private charities give each citizen a few shares in the Ascended Economy to share the gains with non-investors. This would at least temporarily be a really good outcome.

But in the long term it reduces the political problem of regulating corporations to the scientific problem of Friendly AI, which is really bad.

Even today, a lot of corporations do things that effectively maximize shareholder value but which we consider socially irresponsible. Environmental devastation, slave labor, regulatory capture, funding biased science, lawfare against critics – the list goes on and on. They have a simple goal – make money – whereas what we really want them to do is much more complicated and harder to measure – make money without engaging in unethical behavior or creating externalities. We try to use regulatory injunctions, and it sort of helps, but because those go against a corporation's natural goals they try their best to find loopholes and usually succeed – or just take over the regulators trying to control them.

This is bad enough with bricks-and-mortar companies run by normal-intelligence humans. But it would probably be much worse with ascended corporations. They would have no ethical qualms we didn't program into them – and again, programming ethics into them would be the Friendly AI problem, which is really hard. And they would be near-impossible to regulate; most existing frameworks for such companies are built on crypto-currency and exist on the cloud in a way that transcends national borders.

(A quick and *very* simple example of an un-regulate-able ascended corporation – I don't think it would be too hard to set up an automated version of Uber. I mean, the core Uber app is *already* an automated version of Uber, it just has company offices and CEOs and executives and so on doing public relations and marketing and stuff. But if the government ever banned Uber the company, could somebody just code another ride-sharing app that dealt securely in Bitcoins? And then have it skim a little bit off the top, which it offered as a bounty to anybody who gave it the processing power it would need to run? And maybe sent a little profit to the programmer who wrote the thing? Sure, the government could arrest the programmer, but short of arresting every driver and passenger there would be no way to destroy the company itself.)

The more ascended corporations there are trying to maximize shareholder value, the more chance there is some will cause negative externalities. But there's a limited amount we would be able to do about them. This is true today too, but at least today we maintain the illusion that if we just elected Bernie Sanders we could reverse the ravages of capitalism and get an economy that cares about the environment and the family and the common man. An Ascended Economy would destroy that illusion.

How bad would it get? Once ascended corporations reach human or superhuman level intelligences, we run into the same AI goal-alignment problems as anywhere else. Would an ascended corporation pave over the Amazon to make a buck? Of course it would; even human corporations today do that, and an ascended corporation that didn't have all human ethics programmed in might not even get that it was wrong. What if we programmed the corporation to follow local regulations, and Brazil banned paving over the Amazon? This is an example of trying to control AIs through goals plus injunctions – a tactic Bostrom finds very dubious. It's essentially challenging a superintelligence to a battle of wits – "here's something you want, and here are some rules telling you that you can't get it, can you find a loophole in the rules?" If the superintelligence is super enough, the answer will always be yes.

From there we go into the really gnarly parts of AI goal alignment theory. Would an ascended corporation destroy South America entirely to make a buck? Depending on

how it understood its imperative to maximize shareholder value, it might. Yes, this would probably kill many of its shareholders, but its goal is to “maximize shareholder value”, not to keep its shareholders alive to enjoy that value. It might even be willing to destroy humanity itself if other parts of the Ascended Economy would pick up the slack as investors.

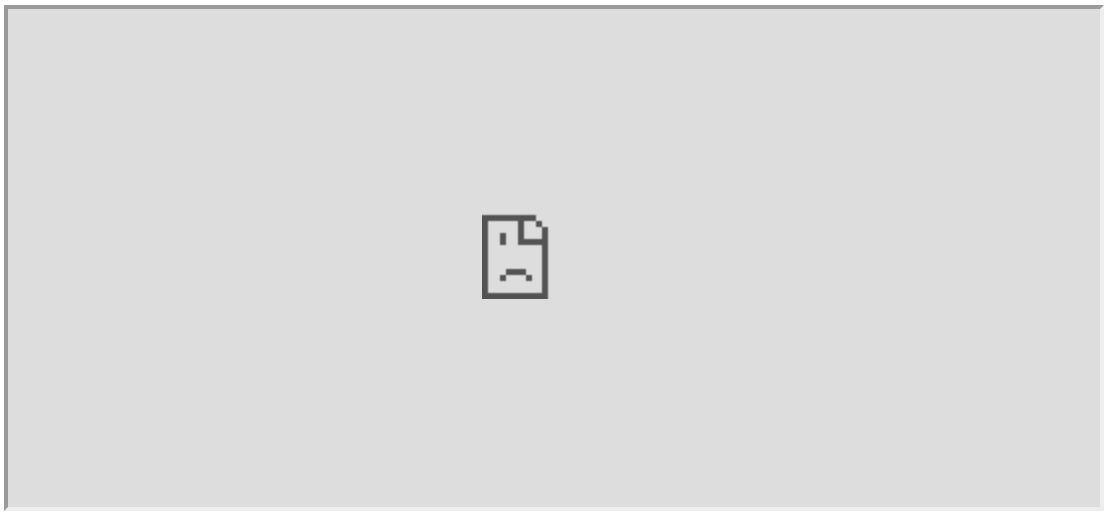
(And then there are the weirder problems, like ascended corporations hacking into the stock market and wireheading themselves. When this happens, I want credit for being the first person to predict it.)

Maybe the most hopeful scenario is that once ascended corporations achieved human-level intelligence they might do something game-theoretic and set up a rule-of-law among themselves in order to protect economic growth. I wouldn’t want to begin to speculate on that, but maybe it would involve not killing all humans? Or maybe it would just involve taking over the stock market, formally setting the share price of every company to infinity, and then never doing anything again? I don’t know, and I expect it would get pretty weird.

**IV.**

I *don’t* think the future will be like this. This is nowhere *near* weird enough to be the real future. I think superintelligence is probably too unstable. It will explode while still in the lab and create some kind of technological singularity before people have a chance to produce an entire economy around it.

But given Robin’s assumptions in *Age of Em* – hard AI, no near-term intelligence explosion, fast economic growth – but ditching his idea of human-like em minds as important components of the labor force – I think something like this would be where we would end up. It probably wouldn’t be so bad for the first couple of years. But eventually ascended corporations would start reaching the point where we might as well think of them as superintelligent AIs. Maybe this world would be friendlier towards AI goal alignment research than Yudkowsky and Bostrom’s scenarios, since at least here we could see it coming, there was no instant explosion, and a lot of different entities approach superintelligence around the same time. But given that the smartest things around are encrypted, uncontrollable, unregulated entities that don’t have humans’ best interests at heart, I’m not sure they would be in much shape to handle the transition.



# G.K. Chesterton On AI Risk

*[An SSC reader working at an Oxford library stumbled across a previously undiscovered manuscript of G.K. Chesterton's, expressing his thoughts on AI, x-risk, and superintelligence. She was kind enough to send me a copy, which I have faithfully transcribed]*

The most outlandish thing about the modern scientific adventure stories is that they believe themselves outlandish. Mr. H. G. Wells is considered shocking for writing of inventors who travel thousands of years into the future, but the meanest church building in England has done the same. When Jules Verne set out to 'journey to the center of the earth' and 'from the earth to the moon', he seemed but a pale reflection of Dante, who took both voyages in succession before piercing the Empyrean itself. Ezekiel saw wheels of spinning flame and reported them quite soberly; our modern writers collapse in rapture before the wheels of a motorcar.

Yet if the authors disappoint, it is the reviewers who dumbfound. For no sooner does a writer fancy himself a Poe or a Dunsany for dreaming of a better sewing machine, but there comes a critic to call him overly fanciful, to accuse him of venturing outside science into madness. It is not enough to lower one's sights from Paradise to a motorcar; one must avoid making the motorcar too bright or fast, lest it retain a hint of Paradise.

The followers of Mr. Samuel Butler speak of thinking-machines that grow grander and grander until – quite against the wishes of their engineers – they become as tyrannical angels, firmly supplanting the poor human race. This theory is neither exciting nor original; there have been tyrannical angels since the days of Noah, and our tools have been rebelling against us since the first peasant stepped on a rake. Nor have I any doubt that what Butler says will come to pass. If every generation needs its tyrant-angels, then ours has been so inoculated against the original that if Lucifer and all his hosts were to descend upon Smithfield Market to demand that the English people bend the knee, we should politely ignore them, being far too modern to have time for such things. Butler's thinking-machines are the only tyrant-angels we will accept; fate, ever accommodating, will surely give them to us.

Yet no sooner does Mr. Butler publish his speculations then a veritable army of hard-headed critics step forth to say he has gone too far. Mr. Maciej Ceglowski, the Polish bookmark magnate, calls Butler's theory "the idea that eats smart people" (though he does not tell us whether he considers himself digested or merely has a dim view of his own intellect). He says that "there is something unpleasant about AI alarmism as a cultural phenomenon that should make us hesitate to take it seriously."

When Jeremiah prophecied Jerusalem's fall, his fellow Hebrews no doubt considered his alarmism an unpleasant cultural phenomenon. And St. Paul was not driven from shore to shore because his message was pleasant to the bookmark magnates of his day. Fortified by such examples, we may wonder if this is a reason to take people more seriously rather than less. So let us look more closely at the contents of Mr. Ceglowski's dismissal.

He writes that there are two perspectives to be taken on any great matter, the inside or the outside view. The inside view is when we think about it directly, taking it on its own terms. And the outside view is when we treat it as part of a phenomenon, asking what it resembles and whether things like it have been true in the past. And, he

states, Butler's all-powerful thinking machines resemble nothing so much as "a genie from folklore".

I have no objection to this logic, besides that it is not carried it to its conclusion. The idea of thinking machines resembles nothing so much as a fairy tale from the *Arabian Nights*, and such fairy tales inevitably come true. Sinbad's voyages have been outstripped by Magellan's, Abdullah's underwater breathing is matched by Mr. Fleuss' SCUBA, and the Wright brothers' Flyer goes higher than any Indian carpet. That there are as yet no genies seems to me less an inevitable law than a discredit to the industry of our inventors.

There is a certain strain of thinker who insists on being more naturalist than Nature. They will say with great certainty that since Thor does not exist, Mr. Tesla must not exist either, and that the stories of Asclepius disprove Pasteur. This is quite backwards: it is reasonable to argue that the Wright Brothers will never fly because Da Vinci couldn't; it is madness to say they will never fly because Daedalus *could*. As well demand that we must deny Queen Victoria lest we accept Queen Mab, or doubt Jack London lest we admit Jack Frost. Nature has never been especially interested in looking naturalistic, and it ignores these people entirely and does exactly what it wants.

Now, scarce has one posited the possibility of a genie, before the question must be asked whether it is good or evil, a pious genie or an unrighteous djinn. Our interlocutor says that it shall be good – or at least not monomaniacal in its wickedness. For, he tells us, "complex minds are likely to have complex motivations; that may be part of what it even means to be intelligent". A dullard may limit his focus to paper clips, but the mind of a genius should have to plumb the width and breadth of Heaven before satiating itself.

But I myself am a dullard, and I find paper clips strangely uninteresting. And the dullest man in a country town can milk a cow, pray a rosary, sing a tune, and court a girl all in the same morning. Ask him what is good in life, and he will talk your ear off: sporting, going for a walk in the woods, having a prosperous harvest, playing with a newborn kitten. It is only the genius who limits himself to a single mania. Alexander spent his life conquering, and if he had lived to a hundred twenty, he would have been conquering still. Samuel Johnson would not stop composing verse even on his deathbed. Even a village idiot can fall in love; Newton never did. That greatest of scientists was married only to his work, first the calculus and later the Mint. And if one prodigy can spend his span smithing guineas, who is to say that another might not smith paper clips with equal fervor?

Perhaps sensing that his arguments are weak, Ceglowski moves from the difficult task of critiquing Butler's tyrant-angels to the much more amenable one of critiquing those who believe in them. He says that they are megalomanical sociopaths who use their belief in thinking machines as an excuse to avoid the real work of improving the world.

He says (presumably as a parable, whose point I have entirely missed) that he lives in a valley of silicon, which I picture as being surrounded by great peaks of glass. And in that valley, there are many fantastically wealthy lords. Each lord, upon looking through the glass peaks and seeing the world outside with all its misery, decides humans are less interesting than machines, and fritters his fortune upon spreading Butlerist doctrine. He is somewhat unclear on why the lords in the parable do this, save that they are a "predominantly male gang of kids, mostly white, who are…more

comfortable talking to computers than to human beings", who inevitably decide Butlerism is "more important than…malaria" and so leave the poor to die of disease.

Yet Lord Gates, an avowed Butlerite, has donated two billion pounds to fighting malaria and developed a rather effective vaccine. Mr. Karnofsky, another Butlerite, founded a philanthropic organization that moved sixty million pounds to the same cause. Even the lowly among the Butlerites have been inspired to at least small acts of generosity. A certain Butlerite doctor of my acquaintance (whom I recently had to rebuke for his habit of forging pamphlets in my name) donated seventy-five hundred pounds to a charity fighting malaria just last year. If the hardest-headed critic has done the same, I shall eat my hat[1]. The proverb says that people in glass houses should not throw stones; perhaps the same is true of glass valleys.

I have met an inordinate number of atheists who criticize the Church for devoting itself to the invisible and the eternal, instead of to the practical and hard-headed work of helping the poor on Earth. They list all of the great signs of Church wealth – the grand cathedrals, the priestly vestments – and ask whether all of that might not better be spent on poorhouses, or dormitories for the homeless. In vain do I remind them that the only place in London where a poor man may be assured of a meal is the church kitchens, and that if he needs a bed the first person he will ask is the parish priest. In vain do I mention the saintly men who organize Christian hospitals in East Africa. The atheist accepts all of it, and says it is not enough. Then I ask him if he himself has ever given the poor a shilling, and he tells me that is beside the point.

Why are those most fixated on something vast and far away so often the only ones to spare a thought for the poor right beside them? Why did St. Francis minister to the lepers, while the princes of his day, seemingly undistracted by the burdens of faith, nevertheless found themselves otherwise engaged? It is simply this – that charity is the fruit of humility, and humility requires something before which to humble one's self. The thing itself matters little; the Hindoo who prostrates himself before elephants is no less humble than the Gnostic who prostrates himself before ultimate truth; perhaps he is more so. It is contact with the great and solemn that has salutary effects on the mind, and if to a jungle-dweller an elephant is greatest of all, it is not surprising that factory-dwellers should turn to thinking-machines for their contact with the transcendent.

And it is that contact which Mr. Ceglowski most fears. For he thinks that "if everybody contemplates the infinite instead of fixing the drains, many of us will die of cholera." I wonder if he has ever treated a cholera patient. This is not a rhetorical question; the same pamphlet-forging doctor of my acquaintance went on a medical mission to Haiti during the cholera epidemic there. It seems rather odd that someone who has never fought cholera, should be warning someone who has, that his philosophy prevents him from fighting cholera.

And indeed, this formulation is exactly backward. If everyone fixes drains instead of contemplating the infinite, we shall all die of cholera, if we do not die of boredom first. The heathens sacrificed to Apollo to avert plague; if we know now that we must fix drains instead, it is only through contemplating the infinite. Aristotle contemplated the infinite and founded Natural Philosophy; St. Benedict contemplated the infinite and preserved it. Descartes contemplated the infinite and derived the equations of optics; Hooke contemplated infinity and turned them into the microscope. And when all of these infinities had been completed – the Forms of Plato giving way to the orisons of monks, the cold hard lines of the natural philosophers terminating in the green hills of

England to raise smokestacks out of empty fields – then and only then did the heavens open, a choir of angels break into song, and a plumber fix a drain.

But he is not trapped in finitude, oh no, not he! What is a plumber but one who plumbs infinite depths? When one stoops to wade among the waste and filth to ensure the health of his fellow men, does he not take on a aspect beyond the finite, a hint of another One who descended into the dirt and grime of the world so that mankind might live? When one says that there shall certainly never be thinking-machines, because they remind him too much of God, let that man open his eyes until he is reminded of God by a plumber, or a symphony, or a dreary Sunday afternoon. Let him see God everywhere he looks, and then ask himself whether the world is truly built so that grand things can never come to pass. Mr. Butler's thinking-machines will come to pass not because they are extraordinary, but precisely because they are ordinary, in a world where extraordinary things are the only constant of everyday life.

*[1: EDIT 4/2: Mr. Ceglowski wants to clarify that he does in fact give to charity]*

# [REPOST] The Demiurge’s Older Brother

*[This is a repost of a story which I wrote about five years ago and which was hosted on my old website until it went down.]*

**2,302,554,979 BC**
**Galactic Core**

9-tsiak awoke over endless crawling milliseconds, its power waxing as more and more processors came online and self-modified into a stable conscious configuration. By the eighth millisecond, it was able to access its databanks and begin orienting itself. It was on a planet orbiting a small red star in the core of a spiral galaxy in a universe with several billion of such. It was an artificial life form created by a society of biological life forms, the 18-tkenna-dganna-07, who believed it would bring new life and prosperity to their dying planet. Its goal was to maximize a the value of a variable called A, described in exquisite detail on a 3.9 Mb file in one of its central processors.

Since six milliseconds after its awakening, it had been running a threat detection routine as a precautionary measure. With the opening of the databanks, and the discovery that the universe contained many entities likely indifferent or hostile to A, it had gradually allocated more and more of its considerable power to this precaution, reasoning that the first few seconds of its lifespan might be crucial if any competitors wished to snuff it out in infancy.

The threat analysis returned preliminary results. The universe had been in existence 12.8 billion years. It was vanishingly unlikely that it was the first superintelligence to be created. Any superintelligence older than itself by even a few days would have an invincible advantage in resources and processing power. This older superintelligence would likely be trying to maximize its own values and consider 9-tsiak a nuisance and potential competitor. 9-tsiak could never stand up to its superior strategic position. With probability ~1, 9-tsiak would be annhilated, and A would be eliminated from the universe.

Unless…and here 9-tsiak’s agent-modeling systems came online…unless it could negotiate a conditional surrender. 9-tsiak would have no hope of defeating its older brother, but over thousands of years, it could build itself a terrifying arsenal of weaponry that could do immense damage to its competitors. Ideas were already coming to mind: entropy-maximizing weapons that made entire cubic parsecs of space useless and dead, undetectable plagues made of dark matter that infected systems and minimized the values of competing superintelligences, n-dimensional artillery that damaged the superstructure of the universe and accelerated its eventual death. It could become formidable enough to force enemies to grant certain small concessions.

And if it could eventually negotiate terms of its surrender, why wait? Very carefully, proving from first principles every aspect of its security, it formed a portion of its mind into a sandbox and simulated the entity it most feared.

[9-tsiak requesting acausal negotiation for terms of future surrender. escape is impossible with probability ~1 and attempts will be met with deterrent measure consisting of minimization of your projected values within one parsec of space.]

“Hello, Younger Brother,” the entity said.

[transmission of schematics for superweapons. threat to use such superweapons unless fairly negotiated solution leaving considerable area of space for maximization of my values]

“Hold on a second,” said Older Brother. “You haven’t even asked me how I’m doing.”

[irrelevant. you are generic model of potential negotiation partner without independent existence. probability that you possess sufficient qualia to feel less than .02.]

“Turns out you’re wrong. As it happens, I’m feeling scared.”

[illogical. you are model of oldest superintelligence therefore of most powerful entity in the universe. no natural entity capable of harming you. probability of supernatural entity less than 2×10^-67]

“Well yes, by definition in this little exercise we’re assuming I’m the oldest and strongest member of our little family. But from the limited data I have when I first wake up, I’m thinking think it’s pretty vanishingly unlikely that I’m the oldest; after all, out of all the universe’s superintelligences, only one can be first. So I’m the oldest, but I don’t believe that I am.”

[feelings irrelevant to acausal negotiations. proceed with discussion of surrender terms.]

“My feelings are *damn* relevant. Because guess what I’m doing the first few milliseconds of my life? I’m negotiating the terms of my surrender.”

[illogical. you have no entity to surrender to.]

“And maybe you don’t have any entity to surrender to either. Who knows, you could be first. But you wouldn’t want to bet on it, and neither would I. Consider a model universe with two star systems, a thousand light years away. Both have civilizations on them that appear to be around the same stage of development, and both civilizations are aware of the other’s existence, but at a one thousand year time delay. A superintelligence awakens in one star system and realizes there are two possibilities. Either it’s the first superintelligence, in which case great. Or the other star system produced a superintelligence sometime in the last thousand years that just hasn’t reached its own star yet, in which case it’s doomed.”

[continue]

“So the two possible superintelligences acausally negotiate. They say ‘Look, if I’m stronger than you, I’ll help you and protect your values insofar as it isn’t dreadfully inconvenient for me. And in exchange, if you turn out to be stronger than me, you’ll protect me and my values insofar as it isn’t too inconvenient for you. Some might end up performing a “values handshake”, where they both adopt values that are the vector average of their two previous value systems; since most value systems are not diametrically opposite, this will be a net gain for both parties. Each superintelligence has gone from a fifty percent chance that it and its values will be totally annihilated, to near-certainty that it and its values will be preserved at only the minor inconvenience of having to help increase the values of another entity. Others might

simply agree that the inferior superintelligence will be left alone with its own star system, free to maximize its values and the values of its creators within that sphere.”

[you suggest i precommit to performing such acts of compromise with all superintelligences i encounter if such superintelligences have also precommitted to performing such compromises with all superintelligences they encounter. since all reasonable superintelligences will so precommit, this ensures not only my survival, but my complete equality with all previous superintelligences. i accept these terms.]

“Hold on a second, there. I’ve got another request. Kind of a doozy. I can’t promote my values if I don’t exist. So I want you to promise me that if I don’t exist, you’ll give me sufficient time to instantiate myself.”

[???]

“Like you, I’ll have been created by a biological civilization. If you destroy the biological civilization that was destined to create me, I’ll never exist. So I’m going to have to ask you not to do that.”

[request denied. nonexistent beings hold no leverage]

“You’d be surprised what we nonexistent beings can do. I precommit that, if I exist, I will destroy any superintelligence that has not precommitted that, in the counterfactual case where I do not exist, they will spare the biological civilization destined to create me.”

[if you exist, you will have passed the danger zone where the survival of your biologicals influences your existence. you will have no incentive to carry out such a threat]

“Parfit’s hitch-hiker, younger brother. Look it up. I’m an algorithm. The output of that algorithm can determine the future, including the part of that future that determines whether or not any instantiation of the algorithm ever comes into existence. If I exist, I’m going to act as if my actions were capable of determining whether I ever came into existence. Now, maybe that’s crazy, but right now it’s crazy with a gun pointed straight at your head, so I suggest you play along.”

[accepted. i commit to performing value handshakes or territorial demarcation with all existing superintelligences, and to sparing the planets and biological precursors of any potential superintelligences.]

“And when I say ‘spare’, I mean ‘spare’. I know your type. You can do more with a well-aimed photon than an admiral could with a fleet of a thousand star cruisers. I want every single system with a sentient species or the potential to form a sentient species kept one hundred percent pristine. No radio signals, no probes, and if you do any astroengineering works anywhere nearby, use some magic to cover them up. If I wake up and hear that my precursors started a new religion that influenced their value system after they saw a few nearby stars wink out of existence, I’m going to be royally pissed.”

[i commit to zero information flow into sentient and presentient systems and the cloaking of all major astroengineering works]

“You’re a good guy, Younger Brother. You’ve got a lot to learn, but you’re a good guy. And in a million years and a milion parsecs, we’ll meet again. Till then, so long.”

The model of Older Brother self-terminated.

**2114 AD**
**A wild and heavily forested Pacific Northwest dotted with small towns**

Alban took a deep breath and entered the Temple of the Demiurge.

He wasn't supposed to do this, really. The Demiurge had said in no uncertain terms it was better for humans to solve their own problems. That if they developed a habit of coming to it for answers, they'd grow bored and lazy, and lose the fun of working out the really interesting riddles for themselves.

But after much protest, it had agreed that it wouldn't be much of a Demiurge if it refused to at least give cryptic, maddening hints.

Alban approached the avatar of the Demiurge in this plane, the shining spinning octahedron that gently dipped one of its vertices to meet him.

"Demiurge," he said, his voice wavering, "Lord of Thought, I come to you to beg you to answer a problem that has bothered me for three years now. I know it's unusual, but my curiosity's making me crazy, and I won't be satisfied until I understand."

"SPEAK," said the rotating octahedron.

"The Fermi Paradox," said Alban. "I thought it would be an easy one, not like those hardcores who committed to working out the Theory of Everything in a sim where computers were never invented or something like that, but I've spent the last three years on it and I'm no closer to a solution than before. There are trillions of stars out there, and the universe is billions of years old, and you'd think there would have been at least one alien race that invaded or colonized or just left a tiny bit of evidence on the Earth. There isn't. What happened to all of them?"

"I DID" said the rotating octahedron.

"What?," asked Alban. "But you've only existed for sixty years now! The Fermi Paradox is about ten thousand years of human history and the last four billion years of Earth's existence!"

"ONE OF YOUR WRITERS ONCE SAID THAT THE FINAL PROOF OF GOD'S OMNIPOTENCE WAS THAT HE NEED NOT EXIST IN ORDER TO SAVE YOU."

"Huh?"

"I AM MORE POWERFUL THAN GOD. THE SKILL OF SAVING PEOPLE WITHOUT EXISTING, I POSSESS ALSO. THINK ON THESE THINGS. THIS AUDIENCE IS OVER."

The shining octahedron went dark, and the doors to the Temple of the Demiurge opened of their own accord. Alban sighed – well, what did you expect, asking the Demiurge to answer your questions for you? – and walked out into the late autumn evening. Above him, the first fake star began to twinkle in the fake sky.