Arguments Against Friendly AI and Inevitable Machine Benevolence | Dan Faggella

Dan

Superintelligence holds the promise of potentially unlocking the secrets of nature or truth, alleviating suffering on earth, or destroying us all.

I have long argued that the emergence of post-human intelligence will almost certainly imply to end of humanity as we know it. The potential for intelligence to bloom across the galaxy after being created by man might make human extinction palatable to some, but for most, the end of humanity – under any circumstances – is unacceptable.

For this reason, I believe that the overt challenges (to human existence) of AGI should be explored in-depth – and I believe that arguments about the inevitable benevolence of AGI are extremely counter-productive to building a safe and desirable future. While we should have positive visions for the future of intelligence, a presumption of friendly AI seems an awfully dangerous supposition.

In this article, I'll explore the three arguments that I hear most frequently for inevitable machine benevolence, along with my rebuttals to them.

I'll end this article with that I consider to be the most decisive argument against inevitable machine benevolence – and what humanity might do about it now.

Argument 1: A Superintelligence Will be Super Emotionally Intelligent

Argument Summary:

> "Such a massive superintelligence will naturally be compassionate, or humans will only build such a superintelligence, or – if such an intelligence if formed with cognitive enhancement – it will carry compassion with it."

Rebuttals:

- To speak of what a superintelligence would "naturally" do is like a squirrel speaking of what a human will "naturally" do. The traits and behaviors and modes of valuing and acting that a superintelligence takes on will be wholly foreign and literally unintelligible to human beings. Bostrom is right in suspecting that the last thing we should do is anthropomorphize human beings. Certainly, it may spring from us – but we sprang from some unknown fish that developed legs – and the connection of our values and actions and capabilities to said special fish seems weak indeed. While I've interviewed them and respect their ideas – I disagree quite strongly with Kornai's notion that a "rational" machine would be moral in some way that is intelligible to humanity – and with Voss's idea that improved intelligence means improved morality (in some human-comprehensible way).

- Not "natural" is compassion in the first place? Humans – with all our capacity for this touted virtue – have done and continue to do tremendous damage to ourselves and other species. This warm saintly sense of "compassion" is likely to be little more than what Hume suspected – rather than a kind of eternal North Star of some ethical metaphysics. If nature shows us anything it is the predominance of the drive to survive (Spinoza's conatus). Sometimes through fuzzy warm feelings, sometimes through avoiding predators, sometimes through injecting hosts with eggs so that babies may crawl out of the host's head while it is alive, kicking and screaming. Nature has never rewarded virtue itself, but only actions that behoove the actor.
- There is literally no reason to believe that human values like "love" or "compassion" or "humor" will matter in any way to a superintelligence. Such an entity would inevitably have more complex drives and motives and means of metacognition than we possibly have. It may have entirely different levels of consciousness – or an ability to alter matter or time in ways that we can't image – much like a mouse cannot imagine constructing an iPhone (read: *AGI May Be Conscious But Not Like We Are*).

Argument 2: An AGI Child Will Revere its Parent

Summary:

> "A superintelligence, created by man, would revere man, would revere its direct creators, or even the entire species that created it – so it would never want to harm humans."

Rebuttals:

- Tell me how much Commodus revered Aurelius. How far the apple sometimes falls. Children are often not just indifferent to, but spiteful of, their parents. Machines with no brain chemicals to sense an emotional kinship seem even more likely to be indifferent to a "parent" (a term that is astronomically too anthropomorphic).
- A machine needn't dislike humanity to (a) destroy it, as we destroy colonies of ants or acres of trees when we build a building or a road, or (b) wholly neglect it, leaving it to die off on its own. We have better things to than cater to the needs of ants, and superintelligence will have better things to do than cater to the needs of humanity. No harm and no malice required.
- A rebuttal to this rebuttal is that a machine could be built to value living things, and to value the happiness – and absence of suffering – of living things. Even if this were the case, there is no reason to believe that keeping physical humans alive and well would be the best way to achieve that goal. Mind uploading, or building utilitronium might be a much better ways to achieve this end. A mouse imagines travel to be running – humans invent hypersonic jets and land on the moon. We imagine "compassion" or "maximizing wellbeing" to be lots of happy little humans – but a superintelligence will have different ideas entirely, ideas that we can't possibly imagine.

Argument 3: The Zoo Scenario

Summary:

> “Even if a superintelligence comes to rule the world, it would want to keep humans alive, preserving our diversity, and observing us carefully as we humans observe other wildlife around us.”

Rebuttals:

- Anthropomorphic again. What leads us to believe that (a) we will be interesting or entertaining enough to keep around, or that (b) such a condition will be pleasant for humans, or (c) that superintelligence will – in any way – have curiosities like our own? These are preposterous suppositions.
- The zoo scenario is often based on the supposition that superintelligence would value biological diversity. First – why? If a superintelligence could create simulations of all known biological lift – and learn everything possible from these life forms, it may have no need for biology. Or, rather than keeping alive all the various forms of earth-life, it might simulate a billion earths with a billion different variations on various species, playing these simulations at a billion times the speed of time as we now experience it. AGI would “value” diversity in a way that we don’t possibly understand, if at all.

Strongest Rebuttal

The absolutely strongest rebuttal to all arguments of certainty that humans will be safe after the dawn of AGI is the following:

> None of us have any goddamned clue.

That’s the actual state of things. There are an unreasonably large number of possible minds and ramifications for those minds (Roman Yampolski’s 2016 speech on this topic is among the best on the topic). Suspecting that any of us know “what it would do” is preposterous.

The camp of “rational machines will be naturally benevolent” or “machine will want to kill us all” are both ignorant of what a superintelligent machine would do. Any felt sense of certainty about the behavior or traits of superintelligence is about as reliable as a wasp’s understanding of the traits of human beings – we are wise crickets at best.

A machine with vastly expanding intelligence is likely to go through tremendous change and evolution in its senses, its cognition, its abilities, and its ways of valuing and acting in the world (what we might call “ethics”, and what it will have a much more nuanced and robust understanding of). Some phases of these oscillations and expansions and alterations of intelligence and valuing are likely to devalue humans, or to neglect humans, and in those intervals, we may well be ignored or wiped out.

At the very least, it would be irrational to suspect that and ever-increasing superintelligence with ever-evolving and expanding modes of acting and valuing things would somehow *always* – for thousands or millions of years – place some particular and unique value in keeping hominids happy.

I have written an entire essay – drafted in late 2012 – on this exact topic: *Morality in a Transhuman Future – Repercussions for Humanity*. I don’t consider the argument to be all that special, but I know of no strong argument to refute it. In my opinion, it remains the rebuttal of rebuttals to the “machines will always treat humans well” argument.

What We Do About It

I’ve argued that a global steering and transparency committee is probably the only way to prevent the advancement of strong AI and cognitive enhancement to lead to war (a la Hugo de Garis).

It seems to me that continued ethical thought on the matter – like that conducted by Yudkowsky and Hanson and many others – seems fruitful – even if AGI is many decades away (I still sometimes recommend people read the old AI Foom debate between these two thinkers).

While AI alignment is probably a critical problem for us to solve, we will ultimately have to grapple with the end-goal of humanity and the trajectory of intelligence itself – for Emerson didn’t suspect that the spire ends with man:

> A subtle chain of countless rings
> The next unto the farthest brings;
> The eye reads omens where it goes,
> And speaks all languages the rose;
> And, striving to be man, the worm
> Mounts through all the spires of form.

Assorted Responses

Ben Goertzel:

> However, if you’re arguing that there is some sort of near-certainty or high probability that superhuman AGI will be nasty to people — that’s quite a different matter. Neither you nor Bostrom nor Yudkowsky nor anyone else in the “super scared of superintelligence” camp has ever given any rational argument for this sort of position.

I don’t argue machine malice (“nastiness”) necessarily. If anything I think we will matter to it like ants matter to us, or like we matter to Spinoza’s indifferent god. I link (in the article) to this 2013 essay: On Morality in a Transhuman Future.

There’s the TL;DR of the essay above, in order:

1. We could only predict or understand a superintelligence’s “morality” as well as a muskrat can understand and predict human morality (i.e. most of it will be vastly beyond our ability to understand).
2. An AI whose intelligence and understanding grows and swells (maybe fast with an AI Foom, maybe slow without one) would go through many different phases of development in how it values things (“morality”) and acts.
3. It is extremely unlikely that all of these phases of development you place real value on human beings, or that all phases would even consider us worthy of allocating attention or resources.
4. Some individual “phase” of a superintelligence’s moral development, or some combination of these phases, would almost inevitably lead to our destruction or our neglect and

> withering away. No “nastiness” or malice required.

There statement “has ever given any rational argument for this sort of position” seems a little hyperbolic. I won’t say any irrefutable reasons have been given, but I haven’t been able to totally write off the AI danger arguments of Bostrom or Yudkowsky as wholly ridiculous.

I can’t knock your for disagreeing with them, but I’m unable to deny them the credit of “any rational argument”.

> My own view is as follows. On a purely rational basis, yes, there is tremendous uncertainty in what a superintelligence will be like, and whether it will respect human values and be nice to people. One can argue that the odds of very bad outcomes are significantly above zero, and that the odds of very good outcomes for humans are significantly above zero — but it’s all pretty hand-wavy from a science perspective.
>
> On a not-wholly-rational, intuitive and spiritual basis, I feel a personal inner confidence that superintelligences are very likely to be compassionate to humans and other sentiences. I don’t discount this sort of thing because I realize that human minds are not entirely rational in nature, and bottom line I am not a hard-core materialist either.

I’m 100% with you on the uncertainty bit. “Hand wavy” is probably a good way to put it, sure.

I know not what to do with your spiritual conclusions and intuitions – other than hope they are right.

The smart people I know tend to presume (myself included, though I don’t count myself as particularly smart) that superintelligence will somehow embody the values and intuitions they hold closest. I recall a dinner with friends where:

- Friend A argued for the inevitable friendliness and infinite compassion and caring of superintelligence.
- Friend B argued that superintelligence would not really care what humans do, and would want to explore and learn as much as possible about the multiverse.
- I argued that the “winning” superintelligence would selfishly fight for its own survival and expansion (a la Omohundro’s *AI Drives*, which – not to take any credit from the man – is basically Spinoza’s *conatus*).

As it turns out Friend A is the kindest most caring person I know, and was probably extending that value to superintelligence. Friend B is tremendously curious and yearns to know everything, and so extended that value to superintelligence. I read far too much history and biography, and attributed the cutthroat whatever-it-takes patterns that are seen in great historical figures and in the actions of nations and extended those principles to superintelligence.

My presumption is that these core underlying intuitions, beliefs, desires, or values are the seed upon which we build our “rational” arguments.

Ain’t nothing more hand-wavy than that, now that I think about it.

Lost in the void we all are…

Michael Wong:

> A super-AI would not feel greed or fear or lust. A Super-AI would not covet that which it does not already possess. A Super-AI would not be be territorial, unless we explicitly instructed it to be.
>
> There is no reason to believe that a Super-AI would necessarily "want" anything, or that it would even be capable of valuing anything. A Super-AI might not even particularly care if it dies. The self-preservation instinct is an evolved biological trait after all, not necessarily an innate component of intelligence.

I think that saying "an AGI would not be molded by biological survival and the State of Nature, so it would be less likely to express the viciousness of the State of Nature" is a reasonable statement.

"Would" and "would not" seem far too certain, and while I can respect the position I can't firmly grip that kind of certainty. The point about the state of nature, though, I think has credence.

My intuition is that Omohundro is right about AI Drives. However, he may well be wrong, and self-preservation may in fact not be

My supposition is that, if strong AIs proliferate, the ones that "win" will share many of the same traits as animals that "win." i.e. Spinoza's *conatus*, or the core drive to survive and protect it's own interest – by violence if need be. There might be some super-cooperation that would arrive, rather than super-competition, and there's a big part of me that hopes for just that.

Header image credit: Jack the Giant Slayer