François Chollet Follow
Nov 27 · 18 min read

The impossibility of intelligence explosion

Transcendence (2014 science-fiction movie)

In 1965, I. J. Good described for the first time the notion of "intelligence explosion", as it relates to artificial intelligence (AI):

> *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, provided that the machine is docile enough to tell us how to keep it under control.*

Decades later, the concept of an "intelligence explosion"—leading to the sudden rise of "superintelligence" and the accidental end of the human race—has taken hold in the AI community. Famous business leaders are casting it as a major risk, greater than nuclear war or climate change. Average graduate students in machine learning are endorsing it. In a 2015 email survey targeting AI researchers, 29% of respondents answered that intelligence explosion was "likely" or "highly likely". A further 21% considered it a serious possibility.

The basic premise is that, in the near future, a first "seed AI" will be created, with general problem-solving abilities slightly surpassing that of humans. This seed AI would start designing better AIs, initiating a recursive self-improvement loop that would immediately leave human intelligence in the dust, overtaking it by orders of magnitude in a short time. Proponents of this theory also regard intelligence as a kind of superpower, conferring its holders with almost supernatural capabilities to shape their environment—as seen in the science-fiction movie Transcendence (2014), for instance. Superintelligence would thus imply near-omnipotence, and would pose an existential threat to humanity.

This science-fiction narrative contributes to the dangerously misleading public debate that is ongoing about the risks of AI and the need for AI regulation. In this post, I argue that intelligence explosion is impossible—that the notion of intelligence explosion comes from a profound misunderstanding of both the nature of intelligence and the behavior of recursively self-augmenting systems. I attempt to base my points on concrete observations about intelligent systems and recursive systems.

A flawed reasoning that stems from a misunderstanding of intelligence

The reasoning behind intelligence explosion, like many of the early theories about AI that arose in the 1960s and 1970s, is sophistic: it considers "intelligence" in a completely abstract way, disconnected from its context, and ignores available evidence about both intelligent systems and recursively self-improving systems. It doesn't have to be that way. We are, after all, on a planet that is literally packed with intelligent systems (including us) and self-improving systems, so we can simply observe them and learn from them to answer the questions at hand, instead of coming up with evidence-free circular reasonings.

To talk about intelligence and its possible self-improving properties, we should first introduce necessary background and context. What are we talking about when we talk about intelligence? Precisely defining intelligence is in itself a challenge. The intelligence explosion narrative equates intelligence with *the general problem-solving ability displayed by individual intelligent agents*—by current human brains, or future electronic brains. This is not quite the full picture, so let's use this definition as a starting point, and expand on it.

Intelligence is situational

The first issue I see with the intelligence explosion theory is a failure to recognize that intelligence is necessarily part of a broader system—a vision of intelligence as a "brain in jar" that can be made arbitrarily intelligent independently of its situation. A brain is just a piece of biological tissue, there is nothing intrinsically intelligent about it. Beyond your brain, your body and senses—your sensorimotor affordances—are a fundamental part of your mind. Your environment is a fundamental part of your mind. Human culture is a fundamental part of your mind. These are, after all, where all of your thoughts come from. You cannot dissociate intelligence from the context in which it expresses itself.

In particular, there is no such thing as "general" intelligence. On an abstract level, we know this for a fact via the "no free lunch" theorem—stating that no problem-solving algorithm can outperform random chance across *all* possible problems. If intelligence is a problem-solving algorithm, then it can only be understood with respect to a *specific* problem. In a more concrete way, we can observe this empirically in that all intelligent systems we know are highly specialized. The intelligence of the AIs we build today is hyper specialized in extremely narrow tasks—like playing Go, or classifying images into 10,000 known categories. The intelligence of an octopus is specialized in the problem of being an octopus. The intelligence of a human is specialized in the problem of being human.

What would happen if we were to put a freshly-created human brain in the body of an octopus, and let in live at the bottom of the ocean? Would it even learn to use its eight-legged body? Would it survive past a few days? We cannot perform this experiment, but we do know that cognitive development in humans and animals is driven by hardcoded, innate dynamics. Human babies are born with an advanced set of reflex behaviors and innate learning templates that drive their early sensorimotor development, and that are fundamentally intertwined with the structure of the human sensorimotor space. The brain has hardcoded conceptions of having a body with hands that can grab, a mouth that can suck, eyes mounted on a moving head that can be used to visually follow objects (the vestibulo-ocular reflex), and these preconceptions are required for human intelligence to start taking control of the human body. It has even been convincingly argued, for

instance by Chomsky, that very high-level human cognitive features, such as our ability to develop language, are innate.

Similarly, one can imagine that the octopus has its own set of hardcoded cognitive primitives required in order to learn how to use an octopus body and survive in its octopus environment. The brain of a human is hyper specialized in the human condition—an innate specialization extending possibly as far as social behaviors, language, and common sense—and the brain of an octopus would likewise be hyper specialized in octopus behaviors. A human baby brain properly grafted in an octopus body would most likely fail to adequately take control of its unique sensorimotor space, and would quickly die off. Not so smart now, Mr. Superior Brain.

What would happen if we were to put a human—brain and body—into an environment that does not feature human culture as we know it? Would Mowgli the man-cub, raised by a pack of wolves, grow up to outsmart his canine siblings? To be smart like us? And if we swapped baby Mowgli with baby Einstein, would he eventually educate himself into developing grand theories of the universe? Empirical evidence is relatively scarce, but from what we know, children that grow up outside of the nurturing environment of human culture don't develop any human intelligence. Feral children raised in the wild from their earliest years become effectively animals, and can no longer acquire human behaviors or language when returning to civilization. Saturday Mthiyane, raised by monkeys in South Africa and found at five, kept behaving like a monkey into adulthood—jumping and walking on all four, incapable of language, and refusing to eat cooked food. Feral children who have human contact for at least some of their most formative years tend to have slightly better luck with reeducation, although they rarely graduate to fully-functioning humans.

If intelligence is fundamentally linked to specific sensorimotor modalities, a specific environment, a specific upbringing, and a specific problem to solve, ***then you cannot hope to arbitrarily increase the intelligence of an agent merely by tuning its brain—no more than you can increase the throughput of a factory line by speeding up the conveyor belt.*** Intelligence expansion can only come from a co-evolution of the mind, its sensorimotor modalities, and its environment. If the gears of your brain were the defining factor of your problem-solving ability, then those rare humans with IQs far outside

the normal range of human intelligence would live lives far outside the scope of normal lives, would solve problems previously thought unsolvable, and would take over the world—just as some people fear smarter-than-human AI will do. In practice, geniuses with exceptional cognitive abilities usually live overwhelmingly banal lives, and very few of them accomplish anything of note. In Terman's landmark "*Genetic Studies of Genius",* he notes that most of his exceptionally gifted subjects would pursue occupations *"as humble as those of policeman, seaman, typist and filing clerk"*. There are currently about seven million people with IQs higher than 150—better cognitive ability than 99.9% of humanity—and mostly, these are not the people you read about in the news. Of the people who have actually attempted to take over the world, hardly any seem to have had an exceptional intelligence; anecdotally, Hitler was a high-school dropout, who failed to get into the Vienna Academy of Art—twice.

People who do end up making breakthroughs on hard problems do so through a combination of circumstances, character, education, intelligence, and they make their breakthroughs through incremental improvement over the work of their predecessors. *Success—expressed intelligence—is sufficient ability meeting a great problem at the right time.* Most of these remarkable problem-solvers are not even *that* clever—their skills seem to be specialized in a given field and they typically do not display greater-than-average abilities outside of their own domain. Some people achieve more because they were better team players, or had more grit and work ethic, or greater imagination. Some just happened to have lived in the right context, to have the right conversation at the right time. Intelligence is fundamentally situational.

Our environment puts a hard limit on our individual intelligence

Intelligence is not a superpower; exceptional intelligence does not, on its own, confer you with proportionally exceptional power over your circumstances. However, it is a well-documented fact that raw cognitive ability—as measured by IQ, which may be debatable—correlates with social attainment for slices of the spectrum that are close to the mean. This was first evidenced in Terman's study, and later confirmed by others—for instance, an extensive 2006 metastudy by Strenze found a visible, if somewhat weak, correlation between IQ and socioeconomic

success. So, a person with an IQ of 130 is statistically far more likely to succeed in navigating the problem of life than a person with an IQ of 70—although this is never guaranteed at the individual level—but here's the thing: this correlation breaks down after a certain point. There is no evidence that a person with an IQ of 170 is in any way more likely to achieve a greater impact in their field than a person with an IQ of 130. In fact, many of the most impactful scientists tend to have had IQs in the 120s or 130s—Feynman reported 126, James Watson, co-discoverer of DNA, 124—which is exactly the same range as legions of mediocre scientists. At the same time, of the roughly 50,000 humans alive today who have astounding IQs of 170 or higher, how many will solve any problem a tenth as significant as Professor Watson?

Why would the real-world utility of raw cognitive ability stall past a certain threshold? This points to a very intuitive fact: that high attainment requires *sufficient* cognitive ability, but that the current *bottleneck* to problem-solving, to *expressed intelligence,* is not latent cognitive ability itself. The bottleneck is our circumstances. Our environment, which determines how our intelligence manifests itself, puts a hard limit on what we can do with our brains—on how intelligent we can grow up to be, on how effectively we can leverage the intelligence that we develop, on what problems we can solve. All evidence points to the fact that our current environment, much like past environments over the previous 200,000 years of human history and prehistory, does not allow high-intelligence individuals to fully develop and utilize their cognitive potential. A high-potential human 10,000 years ago would have been raised in a low-complexity environment, likely speaking a single language with fewer than 5,000 words, would never have been taught to read or write, would have been exposed to a limited amount of knowledge and to few cognitive challenges. The situation is a bit better for most contemporary humans, but there is no indication that our environmental opportunities currently outpace our cognitive potential.

> *"I am, somehow, less interested in the weight and convolutions of Einstein's brain than in the near certainty that people of equal talent have lived and died in cotton fields and sweatshops."—Stephen Jay Gould*

A smart human raised in the jungle is but a hairless ape. Similarly, an AI with a superhuman brain, dropped into a human body in our modern world, would likely not develop greater capabilities than a smart

contemporary human. If it could, then exceptionally high-IQ humans would already be displaying proportionally exceptional levels of personal attainment; they would achieve exceptional levels of control over their environment, and solve major outstanding problems— which they don't in practice.

Most of our intelligence is not in our brain, it is externalized as our civilization

It's not just that our bodies, senses, and environment determine how much intelligence our brains can develop—crucially, our biological brains are just a small part of our whole intelligence. Cognitive prosthetics surround us, plugging into our brain and extending its problem-solving capabilities. Your smartphone. Your laptop. Google search. The cognitive tools your were gifted in school. Books. Other people. Mathematical notation. Programing. The most fundamental of all cognitive prosthetics is of course language itself—essentially an operating system for cognition, without which we couldn't think very far. These things are not merely *knowledge* to be fed to the brain and used by it, they are literally *external cognitive processes*, non-biological ways to run threads of thought and problem-solving algorithms—across time, space, and importantly, across individuality. These cognitive prosthetics, not our brains, are where most of our cognitive abilities reside.

We are our tools. An individual human is pretty much useless on its own—again, humans are just bipedal apes. It's a collective accumulation of knowledge and external systems over thousands of years—what we call "civilization"—that has elevated us above our animal nature. When a scientist makes a breakthrough, the thought processes they are running in their brain are just a small part of the equation—the researcher offloads large extents of the problem-solving process to computers, to other researchers, to paper notes, to mathematical notation, etc. And they are only able to succeed because they are standing on the shoulder of giants—their own work is but one last subroutine in a problem-solving process that spans decades and thousands of individuals. Their own individual cognitive work may not be much more significant to the whole process than the work of a single transistor on a chip.

An individual brain cannot implement recursive intelligence augmentation

An overwhelming amount of evidence points to this simple fact: a single human brain, on its own, is not capable of designing a greater intelligence than itself. This is a purely empirical statement: out of billions of human brains that have come and gone, none has done so. Clearly, the intelligence of a single human, over a single lifetime, cannot design intelligence, or else, over billions of trials, it would have already occurred.

However, these billions of brains, accumulating knowledge and developing external intelligent processes over thousand of years, implement a system—civilization—which may eventually lead to artificial brains with greater intelligence than that of a single human. It is civilization as a whole that will create superhuman AI, not you, nor me, nor any individual. A process involving countless humans, over timescales we can barely comprehend. A process involving far more *externalized intelligence*—books, computers, mathematics, science, the internet—than *biological intelligence*. On an individual level, we are but vectors of civilization, building upon previous work and passing on our findings. We are the momentary transistors on which the problem-solving algorithm of civilization runs.

Will the superhuman AIs of the future, developed collectively over centuries, have the capability to develop AI greater than themselves? No, no more than any of us can. Answering "yes" would fly in the face of everything we know—again, remember that no human, nor any intelligent entity that we know of, has ever designed anything smarter than itself. What we do is, gradually, collectively, build external problem-solving systems that are greater than ourselves.

However, future AIs, much like humans and the other intelligent systems we've produced so far, will contribute to our civilization, and our civilization, in turn, will use them to keep expanding the capabilities of the AIs it produces. AI, in this sense, is no different than computers, or books, or language itself: it's a technology that empowers our civilization. The advent of superhuman AI will thus be no more of a singularity than the advent of computers, or books, or language. Civilization will develop AI, and just march on. Civilization will eventually transcend what we are now, much like it has

transcended what we were 10,000 years ago. It's a gradual process, not a sudden shift.

The basic premise of intelligence explosion—that a "seed AI" will arise, with greater-than-human problem solving ability, leading to a sudden, recursive, runaway intelligence improvement loop—is false. Our problem-solving abilities (in particular, our ability to design AI) are already constantly improving, because these abilities do not reside primarily in our biological brains, but in our external, collective tools. The recursive loop has been in action for a long time, and the rise of "better brains" will not qualitatively affect it—no more than any previous intelligence-enhancing technology. Our brains themselves were never a significant bottleneck in the AI-design process.

In this case, you may ask, isn't civilization itself the runaway self-improving brain? Is our civilizational intelligence exploding? No. Crucially, the civilization-level intelligence-improving loop has only resulted in measurably *linear* progress in our problem-solving abilities over time. Not an explosion. But why? Wouldn't recursively improving X mathematically result in X growing exponentially? No—in short, because **no complex real-world system can be modeled as `X(t + 1) = X(t) * a, a > 1`**. No system exists in a vacuum, and especially not intelligence, nor human civilization.

What we know about recursively self-improving systems

We don't have to speculate about whether an "explosion" would happen the moment an intelligent system starts optimizing its own intelligence. As it happens, *most* systems are recursively self-improving. We're surrounded with them. So we know exactly how such systems behave—in a variety of contexts and over a variety of timescales. You are, yourself, a recursively self-improving system: educating yourself makes you smarter, in turn allowing you to educate yourself more efficiently. Likewise, human civilization is recursively self-improving, over a much longer timescale. Mechatronics is recursively self-improving—better manufacturing robots can manufacture better manufacturing robots. Military empires are recursively self-expanding—the larger your empire, the greater your military means to expand it further. Personal investing is recursively self-improving—the more money you have, the more money you can make. Examples abound.

Consider, for instance, software. Writing software obviously empowers software-writing: first, we programmed compilers, that could perform "automated programming", then we used compilers to develop new languages implementing more powerful programming paradigms. We used these languages to develop advanced developer tools—debuggers, IDEs, linters, bug predictors. In the future, software will even write itself.

And what is the end result of this recursively self-improving process? Can you do 2x more with your the software on your computer than you could last year? Will you be able to do 2x more next year? Arguably, the usefulness of software has been improving at a measurably linear pace, while we have invested exponential efforts into producing it. The number of software developers has been booming exponentially for decades, and the number of transistors on which we are running our software has been exploding as well, following Moore's law. Yet, our computers are only incrementally more useful to us than they were in 2012, or 2002, or 1992.

But why? Primarily, because the usefulness of software is fundamentally limited by the *context* of its application—much like intelligence is both defined and limited by the context in which it expresses itself. Software is just one cog in a bigger process—our economies, our lives—just like your brain is just one cog in a bigger process—human culture. This context puts a hard limit on the maximum potential usefulness of software, much like our environment puts a hard limit on how intelligent any individual can be—even if gifted with a superhuman brain.

Beyond contextual hard limits, even if one part of a system has the ability to recursively self-improve, other parts of the system will inevitably start acting as bottlenecks. Antagonistic processes will arise in response to recursive self-improvement and squash it—in software, this would be resource consumption, feature creep, UX issues. When it comes to personal investing, your own rate of spending is one such antagonistic process—the more money you have, the more money you spend. When it comes to intelligence, inter-system communication arises as a brake on any improvement of underlying modules—a brain with smarter parts will have more trouble coordinating them; a society with smarter individuals will need to invest far more in networking and communication, etc. It is perhaps not a coincidence that very high-IQ

people are more likely to suffer from certain mental illnesses. It is also perhaps not random happenstance that military empires of the past have ended up collapsing after surpassing a certain size. Exponential progress, meet exponential friction.

One specific example that is worth paying attention to is that of scientific progress, because it is conceptually very close to intelligence itself—science, as a problem-solving system, is very close to being a runaway superhuman AI. Science is, of course, a recursively self-improving system, because scientific progress results in the development of tools that empower science—whether lab hardware (e.g. quantum physics led to lasers, which enabled a wealth of new quantum physics experiments), conceptual tools (e.g. a new theorem, a new theory), cognitive tools (e.g. mathematical notation), software tools, communications protocols that enable scientists to better collaborate (e.g. the Internet)…

Yet, modern scientific progress is measurably linear. I wrote about this phenomenon at length in a 2012 essay titled "*The Singularity is not coming*". We didn't make greater progress in physics over the 1950–2000 period than we did over 1900–1950—we did, arguably, about as well. Mathematics is not advancing significantly faster today than it did in 1920. Medical science has been making linear progress on essentially all of its metrics, for decades. And this is despite us investing exponential efforts into science—the headcount of researchers doubles roughly once every 15 to 20 years, and these researchers are using exponentially faster computers to improve their productivity.

How comes? What bottlenecks and adversarial counter-reactions are slowing down recursive self-improvement in science? So many, I can't even count them. Here are a few. Importantly, every single one of them would also apply to recursively self-improving AIs.

- Doing science in a given field gets exponentially harder over time —the founders of the field reap most the low-hanging fruit, and achieving comparable impact later requires exponentially more effort. No researcher will ever achieve comparable progress in information theory as Shannon did in his 1948 paper.
- Sharing and cooperation between researchers gets exponentially more difficult as a field grows larger. It gets increasingly harder to

keep up with the firehose of new publications. Remember that a network with **N** nodes has **N * (N - 1) / 2** edges.

- As scientific knowledge expands, the time and effort that have to be invested in education and training grows, and the field of inquiry of individual researchers gets increasingly narrow.

In practice, system bottlenecks, diminishing returns, and adversarial reactions end up squashing recursive self-improvement in all of the recursive processes that surround us. Self-improvement does indeed lead to progress, but that progress tends to be linear, or at best, sigmoidal. Your first "seed dollar" invested will not typically lead to a "wealth explosion"; instead, a balance between investment returns and growing spending will usually lead to a roughly linear growth of your savings over time. And that's for a system that is orders of magnitude simpler than a self-improving mind.

Likewise, the first superhuman AI will just be another step on a visibly linear ladder of progress, that we started climbing long ago.

Conclusions

The expansion of intelligence can only come from a co-evolution of brains (biological or digital), sensorimotor affordances, environment, and culture—not from merely tuning the gears of some brain in a jar, in isolation. Such a co-evolution has already been happening for eons, and will continue as intelligence moves to an increasingly digital substrate. No "intelligence explosion" will occur, as this process advances at a roughly linear pace.

Remember:

- Intelligence is situational—there is no such thing as general intelligence. Your brain is one piece in a broader system which includes your body, your environment, other humans, and culture as a whole.

- No system exists in a vacuum; any individual intelligence will always be both defined and limited by the context of its existence, by its environment. Currently, our environment, not our brain, is acting as the bottleneck to our intelligence.

- Human intelligence is largely externalized, contained not in our brain but in our civilization. We are our tools—our brains are modules in a cognitive system much larger than ourselves. A system that is already self-improving, and has been for a long time.
- Recursively self-improving systems, because of contingent bottlenecks, diminishing returns, and counter-reactions arising from the broader context in which they exist, cannot achieve exponential progress in practice. Empirically, they tend to display linear or sigmoidal improvement. In particular, this is the case for scientific progress—science being possibly the closest system to a recursively self-improving AI that we can observe.
- Recursive intelligence expansion is already happening—at the level of our civilization. It will keep happening in the age of AI, and it progresses at a roughly linear pace.

@fchollet, November 2017