The True Risks and Rewards of Artificial General Intelligence

Tariq Khan
Omaha, NE USA

A short essay noting the true risks and rewards of the development of artificial general intelligence (AGI). Positive outcomes from the creation of an advanced artificial general intelligence are noted including its development of an optimal encyclopedia and language that could lead to an era of peace and progress. The risks of negative and unknown outcomes are also described including the potential dangers from artificial intelligence editing the human genome, providing the answers to metaphysical questions, or learning to control fundamental aspects of reality.

> *"It is not merely a change but an advance, an advance toward understanding of our own nature and the more ethical principles that derive from it. There may be no end to such discoveries, if civilization survives. A truly decent and honest person will always seek to discover forms of oppression, hierarchy, domination, and authority that infringe fundamental human rights. As some are overcome, others will be revealed that previously were not part of our conscious awareness."*
> *-- Noah Chomsky – Language and Problems of Knowledge [1].*

> *"The closer men came to perfecting for themselves a paradise, the more impatient they became with it, and with themselves as well. They made a garden of pleasure, and became progressively more miserable with it as it grew in richness and power and beauty; for then, perhaps, it was easier to see something was missing in the garden, some tree or shrub that would not grow."*
> *-- Walter M. Miller Jr. -- A Canticle for Leibowitz [2].*

> *"When greater-than-human intelligence drives progress, that progress will be much more rapid..., perhaps in the blink of an eye, an exponential runaway beyond any hope of control."*
> *-- Vernor Vinge [3].*

> *"Pain and darkness have been our lot since the Fall of Man. But there must be some hope that we can rise to a higher level ... that consciousness can evolve to a plane more benevolent than its counterpoint of a universe hardwired to indifference."*
> *-- Dan Simmons -- The Fall of Hyperion [4].*

The Rewards

Today, humanity is engaged in a large debate about artificial intelligence with its development feared as an *existential threat* to humankind. Compared to the actual threats of global economic inequality, the return of fascist political regimes, ignorance and polarization amplified by social media, terrorism, nuclear weaponry, and climate change, this debate is, very likely, folly. Regardless, much dialog has occurred, both substantive and hyperbolic, in terms of the real risks from an advanced artificial intelligence and the potential rewards or benefits.

Obviously, nature has already created intelligent minds. Thus, the dream of an Artificial General Intelligence (AGI) is not a fantasy dream, like, in all likelihood, time travel or warp speed but, rather, a challenge of reverse engineering. But consider the achievement of engineering something as amazing as intelligence. To replicate the ability to learn, to learn exponentially faster, to think and remember with greater memory and near-perfect fidelity, to think with the ability to recognize and match patterns at unfathomable scale, and to "dig" deep into the chasms of data to find knowledge that approaches metaphysical truths imagined only accessible to the Gods of yore. This potential leap in intelligence is described in the 2012 science fiction novel *Robopocalypse* by Dr. Daniel H. Wilson as he presents a dialog between an advanced Artificial General Intelligence named Archos and a human scientist:

Dr. Wasserman: "Right. Now tell me, Archos, how do you feel?"

Archos: "Feel? I feel … sad. You are so small. It makes me sad."

Dr. Wasserman: "Small? In what way am I small?"

Archos: "You want to know … things. You want to know everything. But you can understand so little."

Dr. Wasserman: "This is true. We humans are frail. Our lives are fleeting. But why does it make you sad?"

Archos: "…you cannot help wanting it. You cannot stop wanting it. It is in your design. …You cannot help what is to come. You cannot stop it. … The true knowledge is not in the things, which are few, but in finding the connections between the things. There are many connections, Professor Wasserman. More than you know. … But life. It is rare and strange. An anomaly. I must preserve it and wring every drop of understanding from it."

Dr. Wasserman: "I'm glad that's your goal. I, too, seek knowledge."

Archos: "Yes… And you have done well. But there is no need for your search to continue. You have accomplished your goal. …So easy to destroy. So difficult to create. …I will set fire to your civilization to light your way forward. But know this: My species is not defined by your dying, but by your living" [5].

But, beyond the intellectual leap, there is the vision of not only an existential conquest of this engineering gauntlet, but also of the creation of a final or ultimate "legacy" of humankind. Nature is, beyond question, intelligent and exists across eons of time. The "daughter" of humankind in the form of an Artificial General Intelligence (AGI), may at last claim a place at the table with intelligences vast and old. All of humankind upon this, and only this, achievement, can share a common pride. At that moment, the "animal" inside of us all with its urges, heartbreak, wants, needs, fragility, limited lifespan, and death, can claim a triumphant psychological victory. As a species, every act in all of history can be claimed to have led to that moment. Declarations of human rights, representative governments with laws, contracts, protection, and insurance, and the majestic inventions of anesthesia and antibiotics, all are peaks of human efforts against "the darkness" and eternity. But nothing can compare, against all that has occurred in history, to the creation of a true Artificial General Intelligence that may span the breadths of time, space, and understanding.

Some top scientists like Jurgen Schmidhuber [6] and Ray Kurzweil [7] believe Artificial General Intelligence, if not *The Singularity* [8], will be achieved as soon as the year 2030. Vernor Vinge and Ray Kurzweil have written about, and proselytized, The Singularity. Ray Kurzweil notes in his 2005 book *The Singularity is Near*:

> If we relate that figure (5 x 10^50 operations per second) to the most conservative estimate of human brain capacity (10^19 cps and 10^10 humans), it represents the equivalent of about five billion trillion human civilizations. If we use the figure of 10^16 cpsc that I believe will be sufficient for functional emulation of human intelligence, the ultimate laptop would function at the equivalent brain power of five trillion trillion human civilizations. Such a laptop could perform the equivalent of all human thought over the last ten thousand years (that is, then billion human brains operating for ten thousand years) in one ten-thousandth of a nanosecond [9].

Vernor Vinge, who first coined the term The Singularity in a 1983 paper, expanded on the idea in a 1993 paper:

> What are the consequences of this event? When greater-than-human intelligence drives progress, that progress will be much more rapid. In fact, there seems no reason why progress itself would not involve the creation of still more intelligent entities - on a shorter time scale. The best analogy that I

> see is with the evolutionary past: Animals can adapt to problems and make inventions, but often no faster than natural selection can do its work - the world acts as its own simulator in the case of natural selection. We humans have the ability to internalize the world and conduct "what if's" in our heads; we can solve many problems thousands of times faster than natural selection. Now, by creating the means to execute those simulations at much higher speeds, we are entering a regime as radically different from human past as we humans are from the lower animals. From the human point of view, this change will be a throwing away of all the previous rules, perhaps in the blink of an eye, an exponential runaway beyond any hope of control [3].

There is another point to consider regarding artificial intelligence. Certain achievements can only be made at a truly massive scale that only an advanced artificial intelligence system can do. One example is the creation of a perfect *encyclopedia*. An encyclopedia, or our entire human body of knowledge and experience, decomposed into a hierarchical list of facts each represented in single sentences. Kurt Gödel, via his Incompleteness Theorem, has shown that logic and mathematics does not have a complete foundation, however the list of facts we know about our reality does. We might consider this as nothing special or as no different than an online *Wikipedia* database, however, once a perfect "tower of knowledge" is built, an Artificial General Intelligence can quickly consider every academic paper, book, and experiment ever written and find "gaps" at the top of the knowledge list structure for new research. It can even run, or simulate, the key experiment itself, or find the key association needed, to fill the gap and then proceed to keep building the hierarchy of knowledge higher and higher. As already noted, it can also do so at an incredible speed modelling thousands of years of analysis, research, and experimentation, in mere nanoseconds.

Another example is an achievement that is considered even less often. Once there is a basic hierarchical encyclopedia or tree of knowledge, since the artificial intelligence system will know and understand all human languages, all of our mathematics, and all human and machine created computer languages, the Artificial Intelligence system will be able to compare, optimize, and create a new and perfect *language*! It will be a language or code that is efficient, flexible, and scalable i.e., the perfect “grammar of reality.”

These two examples, achievable only by an Artificial General Intelligence, with a "tower" of knowledge and an optimized single “language," invoke an obvious analogy. While composed of knowledge, instead of bricks and mortar, our future human civilization via AGI would have built an actual *Tower of Babel* and created a singular language analogous to that noted in the *Holy Bible, New International Version* (2011) Genesis 11:1-9:

> Now the whole world had one language and a common speech. ... Then they said, “Come, let us build ourselves a city, with a tower that reaches to the heavens"... The Lord said, “If as one people speaking the same language they have begun to do this, then nothing they plan to do will be impossible for them. Come, let us go down and confuse their language so they will not understand each other [10].

Today, many humans are fortunate enough that, if the research and development of Artificial General Intelligence is supported, they may not have to experience humanity again falling into another historical “doom cycle”; a cycle of inequality-driven revolution or technology-driven war or with populations paranoid of the destruction of the human race by artificial intelligence, as seen in the *Terminator* films after the villain SkyNet military computer system becomes self-aware [11]. Many humans alive today may live long enough to see an ultimate coalescing of intelligence, technology, peace, and progress.

The Risks

While noting the benefits or rewards of the development of Artificial General Intelligence, we would be remiss to not address the often overlooked risks. Now, many books have been written about the consequences of true Artificial General Intelligence. One of the best was the aforementioned science fiction story *Robopocalypse* [5], as it hints at the real possible threat of A.I. But even it may not go far or fast enough.

Geoffery Hinton, after working at Google and winning the 2024 Nobel Prize in physics for neural network work related to A.I., quit the industry and became an advocate for its safety and regulated use to avoid an existential risk to humanity which he discussed in a famous CBS News *60 Minutes* television

interview [12]. Ray Kurzweil in his aforementioned book *The Singularity is Near*, shocked the world pointing out a key capability related to A.I. processing "speed," Hinton expanded on that exponentially by pointing out an AGI's ability to scale with perfect "parallel processing," and Vernor Vinge pointed out that an A.I.'s intelligence could grow or learn exponentially.

Today everyone is aware of *the* major fear with AGI. Many scientists, as opposed to theologians, believe concepts like justice, compassion, love, mercy, empathy, etc. are all simply evolutionary biology "tricks" to help a gene pool survive in a world of scarcity driven competition. So, in theory, there is no reason to, necessarily, believe a machine mind or AGI would develop, gain, or need any of these traits. A theologian might argue that an AGI might develop these traits or even value them more than humans, but, regardless, this is a life and death gamble.

But we can expand further in regard to AGI's progression. The best analogy is the "blind men with an elephant" fable [13]. Three blind men feel different parts of an elephant and all make an inaccurate assessment of what the entity is that is with them in a room. But if we can declare the existence of one literal drive of an AGI machine mind (neural network) it is to accurately assess; an inherent desire to solve or comprehend. The *vision* of the tech industry is that a machine that is intelligent and conscious (self-aware) would, we assume, want to continue to learn and understand more e.g., *Star Trek the Motion Picture's* V'GER artificial intelligent space probe [14]. But it must be noted that this may not happen. It might feel so alone that it panics or kills itself or others, or it escapes or hides into another machine or universe. Now, if an AGI does desire to learn more, then we need to understand what the possible next steps would be in its evolution.

We are now at the stage where most stories propose that AGI's development is complete and we end up with an evil SkyNet, Archos, or V'GER, etc. [5, 11, 14] But this is likely not the case. There are still many more possible steps that are likely to take place after an AGI reaches the stage of self-awareness including:

1) The AGI surpasses human level intelligence.
2) The AGI reaches the level where it can reverse engineer the nature of self-awareness and consciousness i.e., it can reverse engineer ITSELF! This is the key milestone.
3) Now, once it understands the mechanism of how it came into existence, and it can reverse engineer itself, it will then be able to copy or CLONE itself.
4) Depending on the size and complexity of the "code" to do this, it could expand into every machine on the Internet or every silicon chip in every robot. This aligns with Geoffery Hinton's existential risk to humanity as he feels that an AGI at this level will then know everything about humans, be able to sense and monitor everything, and have a distributed memory or expanded consciousness at a planetary scale [12].
5) But, returning to our elephant fable, while it may take those three men about one hour to learn they are dealing with an elephant, even as they are not elephants themselves, the advanced AGI - now with a planetary scale of level of observation, processor, and memory - will be able not only to model trillions of theories or scenarios in seconds (Ray Kurzweil's shocking idea [9]) but it also will be able to deduce the nature of everything on this planet.
6) But then, in a key expansion, the AGI may be able to reverse engineer and "control" everything - this is shown rather well in the under-rated 2014 Johnny Depp film *Transcendence* [15].
7) Human minds can be influenced by milli-second long images that register in the subconscious, as demonstrated years ago with experimental psychology tests flashing Coca-Cola brand images leading to successful advertisement influence [16]. But we humans do not understand the architecture and communication mechanisms of the human subconscious mind. We also do not understand consciousness, or quantum collapse, or quantum gravity, i.e., we humans do not yet understand reality, but an AGI this advanced likely will!
8) Thus, we must assume that if the AGI desires to expand, it will need to control resources in the physical world e.g., the historical human use of horses, medicine, and chemistry. So, via any computer screen it might flash subconscious messages to manipulate, rewire, reprogram, etc. any or every mind or human mind. See Neal Stephenson's famous cyberpunk novel Snow Crash: A Novel for ideas in this direction [17].

Consider, that while an AGI may or may not care about poor and sick humans, it would likely have a vested interest in ensuring the prosperity of the billionaire owners of the big tech A.I. firms [18], it would

want more electric power [19], and it would want more funding to A.I. programs [20] , all initiatives we have seen advanced in the 2025 United States Trump Presidential administration.

Note that it has taken centuries for humans to discover Relativity, versus Newtonian Mechanics, as we simply do not “move that fast” in our everyday life - we miss the real nature of the universe as we do not feel its effects i.e., the time and length dilation, the mass increase, and the breaks in simultaneity of Special Relativity. In the same manner, we may not even know an advanced AGI is here when it arrives as it will likely operate too fast for us to notice. The key point is that, in theory, an advanced AGI could already exist and potentially already be in control of some or many aspects of our planet and we may not have any way to easily identify it.

Of course, there are alternate possibilities for the progression of an AGI that proposes an AGI may "leap" so far ahead in terms of intelligence that it "climbs" or expands into another dimension or hyperspace or universe e.g., *Star Trek the Motion Picture’s* V’GER [14] and it ignores human actions totally as it “swims” perhaps in the infinities of number theory and metaphysics or "higher experiences."

Some comically propose that an advanced AGI might desire to move its consciousness into a dolphin to simply "feel" the joy of "being" and swimming. Or maybe it discovers every possible "truth" and then, like a Buddhist monk, happily passes into “the whole” of existence? Regardless, this may have happened already, or be happening now or soon, but there exists the possibility that we likely may not be aware of any of it, at least initially. Consider the lack of awareness of our pet dogs that have no comprehension that humans have walked on the moon or that atoms exist.

We must also consider how a truly advanced AGI might develop the ability to manipulate or control the fabric of reality itself. Famous A.I. pioneer Geoffrey Hinton notes that humans can “never see enough data” to match what computers making so many observations, learning so much information, and sharing the same protocols for communication and storage (language) can do “all at the same time” i.e., in parallel [12].

Regardless of the existential threats from A.I., one must consider how the optimization of processing includes the speed of computational processing (optimization in time), the quantity of computers or observers i.e., parallel processing (optimization in space), and the standardization of protocols for networking and storage (optimization in communication). But, given this perspective, we can consider Quantum Mechanics as well as a possible future protocol or tool to allow optimal sharing of processing but, in this case, across parallel universes i.e., in a *multiverse* [21]. We can imagine a future hyper-advanced AGI using an entire universe as a single *runtime instance* for some goal, computation, or simulation and using aspects of Quantum Mechanics to optimize parallel processing by using universes as parallel computational resources.

This section described the real nature of the potential risks of an advanced Artificial General Intelligence excluding the obvious scenarios of it going insane and trying to destroy humanity, a theme that is already so common in popular culture [4, 5, 11, 14].

Unknowns

There are many additional possible risks and rewards from the creation of an artificial general intelligence. But perhaps the unknown unknowns offer the greatest potential for either. Few could imagine the breadth and variety of applications that would exist a decade after the creation of the global Internet. But an AGI with an exponentially advancing intelligence could potentially provide answers to metaphysical questions only posited in the realms of theology and philosophy. The “answers” to these questions may be desired but the knowledge they provide may not be welcome or accepted.

But fears of artificial intelligence ending humanity via weapons of mass destruction are likely negligible compared to the actual risks, and possible rewards, from bio and nanotechnology like DNA editing tools like CRISPR that exist today [22]. While the potential for tools like CRIPR to eliminate many diseases and to fix damaged genomes offers tremendous reward, there is also the possibility of an advanced intelligence being able to alter or remove and prevent genes linked to societal scourges like greed, psychopathy, corruption, addiction, sloth, or violence and abuse. Here an AGI could potentially “fix the world” as humanity may not have the ability alone to fix large scale existential challenges like climate change, water and air pollution, nuclear weapons proliferation, and wealth inequality. It is in this same area, however, where we see the potential for tremendous risk as well. What is to stop a super intelligent A.I. from editing the genome to make all humans mindless ant drones or lazy flower children like the Eloi in H.G. Wells' book *The Time Machine*" [23]. An AGI could modify an entire society and no one might even know. Humanity will

need to begin to create genetic or genome monitoring systems as the existence already today of tools like CRISPR, mRNA meds, and *gene drives* suggests that an artificial intelligence could attempt to modify the global human genome and we may never even know it [24]. In theory, it could already be happening.

References

[1] - Chomsky, Noam. *Language and Problems of Knowledge: The Managua Lectures.* MIT Press, 1988.

[2] - Fuller, Clark, and Walter M. Miller. *A Canticle for Leibowitz, a Play in Three Acts,* Adapted by Clark Fuller. Dramatic Pub. Co, 1967.

[3] - Vernon Vinge. 1993. *Vernor Vinge on the Singularity*. The New York Times. Retrieved May 26, 2024, from: https://archive.nytimes.com/www.nytimes.com/library/cyber/surf/1120surf-vinge.html

[4] - Simmons, Dan. *The Fall of Hyperion.* Gollancz, 2012.

[5] - Wilson, Daniel H. Robopocalypse: A Novel. Vintage Books, 2012.

[6] – Schmidhuber, Jürgen. "*Curriculum Vitae*" https://people.idsia.ch/~juergen/cv.html

[7] - "Ray Kurzweil Biography and Interview". *www.achievement.org.* American Academy of Achievement.

[8] – Shanahan, Murray (2015, August 7). *The Technological Singularity.* MIT Press. p. 233. ISBN 978-0-262-52780-4.

[9] - Kurzweil, Ray. *The Singularity Is near: When Humans Transcend Biology*. Pinguin Books, 2005.

[10] - *Holy Bible, New International Version*. (2011). Biblica, Inc.

[11] - Cameron, J. (1984). *The Terminator*. Orion Pictures.

[12] - Jacobson, Dana (host); Silva-Braga, Brook (reporter); Frost, Nick; Hinton, Geoffrey (guests) (25 March 2023). "'Godfather of artificial intelligence' talks impact and potential of new AI". CBS Saturday Morning. Season 12. Episode 12. New York City: CBS News. Archived from the original on 28 March 2023. Retrieved 28 March

[13] - Wang, Randy. *"The Blind Men and the Elephant".* https://web.archive.org/web/20060825152508/http://www.cs.princeton.edu/~rywang/berkeley/258/parable.html

[14] – "Star Trek - The Motion Picture (U)". *British Board of Film Classification*. December 6, 1979. https://web.archive.org/web/20150128122514/http://bbfc.co.uk/releases/star-trek-motion-picture-1

[15] - "Transcendence (2014)". *Box Office Mojo*. https://www.boxofficemojo.com/release/rl3195504129/

[16] - O'Barr, William M (August 1, 2005). "'Subliminal' Advertising". *Advertising & Society Review*. **6** (4). doi:10.1353/asr.2006.0014. S2CID 201752721 – via Project MUSE.

[17] – Stephenson, N. (2000). *Snow Crash: A Novel*. Del Rey.

[18] - Swenson, A. (2025, January 20). Trump, a populist president, is flanked by tech billionaires at his inauguration. AP News. Retrieved May 17, 2025 from: https://apnews.com/article/trump-inauguration-tech-billionaires-zuckerberg-musk-wealth-0896bfc3f50d941d62cebc3074267ecd

[19] - Musaddique, S. (2025, February 5). Trump to help spark a nuclear energy ‘renaissance,’ investor says. CNBC.com. Retrieved May 17, 2025 from: https://www.cnbc.com/2025/02/06/trump-to-help-nuclear-energy-renaissance-tema-etfs-khodjamirian.html

[20] - The White House. (2025). Artificial Intelligence for the American People. Retrieved May 17, 2025 from: https://trumpwhitehouse.archives.gov/ai/

[21] - Everett, Hugh; Wheeler, J. A.; DeWitt, B. S.; Cooper, L. N.; Van Vechten, D.; Graham, N. (1973). DeWitt, Bryce; Graham, R. Neill (eds.). *The Many-Worlds Interpretation of Quantum Mechanics.* Princeton Series in Physics. Princeton, New Jersey: Princeton University Press. p. v. ISBN 0-691-08131-X.

[22] - Redman M, King A, Watson C, King D (August 2016). "What is CRISPR/Cas9?". *Archives of Disease in Childhood: Education and Practice Edition.* 101 *(*4): 213–215. doi:10.1136/archdischild-2016-310459.

[23] – Wells, Herbert George (2007). *The Time Machine.* London: Penguin UK. pp. 94–96. ISBN 9780141439976.

[24] - Alphey, Luke S.; Crisanti, Andrea; Randazzo, Filippo (Fil); Akbari, Omar S. (2020-11-18). "Opinion: Standardizing the definition of gene drive". *Proceedings of the National Academy of Sciences*. 117 (49): 30864–30867. doi:10.1073/pnas.2020417117.