Plausible Scenarios For Artificial Intelligence in Preventing Catastrophes

Given consistent empirical evidence demonstrating that the raising of red flags has resulted in a lower probability of essential actors heeding warnings than causing of unintended consequences, those endowed with a public platform carry a special responsibility, as do system architects. Nowhere is this truer than at the confluence of advanced technology and existential risk to humanity.

As one who has invested a significant portion of my life studying crises and designing methods of prevention, including for many years the use of artificial intelligence (AI), I feel compelled to offer the following definitions of AI with a few of what I consider to be plausible scenarios on how AI could prevent or mitigate catastrophes, as well as brief enlightenment on how we reduce the downside risk to levels similar to other technology systems.

An inherent temptation with considerable incentives exists in the transdisciplinary field of AI for the intellectually curious to expand complexity infinitely, generally leaving the task of pragmatic simplicity to applied R&D, which is the perspective the following definitions and scenarios are offered, primarily for consideration and use in this article.

A Definition of Artificial Intelligence: *Any combination of technologies and methodologies that result in learning and problem solving ability independent of naturally occurring intelligence.*

A Definition of Beneficial AI: *AI that has been programmed to prevent intentional harm and to mitigate unintended consequences within a strictly controlled governance system, which include redundant safety functions, and continuous human oversight (to include a kill switch in high-risk programs).*[1]

A Definition of Augmentation: For the purposes of this article, augmentation is defined not as implants or direct connection to biological neural systems, which is an important rapidly emerging sub-field of AI in biotech, but rather simply *enhancing the quality of human work products and economic productivity with the assistance of AI*.[2]

To aid in this exercise, I have selected quotes from the highly regarded book *Global Catastrophic Risks* (GCR), which consists of 22 chapters by 25 authors in a worthy attempt to provide a comprehensive view.[3] I have also adopted the GCR definition of '*global catastrophic*': *10 million fatalities or 10 trillion dollars*. The following reminders from GCR seem appropriate in a period of intense interest in AI and existential risk:

> *"Even for an honest, truth-seeking, and well-intentioned investigator it is difficult to think and act rationally in regard to global catastrophic risks and existential risks."*
>
> *"Some experts might be tempted by the media attention they could get by playing up the risks. The issue of how much and in which circumstances to trust risk estimates by experts is an important one."*

[1] In the GCR chapter Artificial Intelligence as a 'Positive and Negative Factor in Global Risk', Eliezer Yudkowsky prefers 'Friendly AI'. His complete chapter is on the web here: https://intelligence.org/files/AIPosNegFactor.pdf

[2] During the course of writing this paper Stanford University announced a 100-year study of AI. The white paper by Eric Horvitz can be viewed here: https://stanford.app.box.com/s/266hrhww2l3gjoy9euar

[3] *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Cirkovic, (Oxford University Press, Sep 29, 2011): http://ukcatalogue.oup.com/product/9780199606504.do

Super Volcanoes

A very high probability event that should consume greater attention, super volcano eruptions occur about every 50,000 years. GCR highlights the Toba eruption in Indonesia approximately 75,000 years ago, which may be the closest humanity has come to extinction. The primary risk from super eruptions is airborne particulates that can cause a rapid decline in global temperatures, estimated to have been 5–15 C after the Toba event.[4]

While it appears that a super-eruption contains a low level of existential risk to the human race, a catastrophic event is almost certain and would likely exceed all previous disasters, followed by massive loss of life and economic collapse, reduced only by the level of preventative measures taken in advance. While preparation and alert systems have much improved since my co-workers and I observed the eruption of Mt. St. Helens from Paradise on Mt. Rainier the morning of May 18th, 1980, it is quite clear that the world is simply not prepared for one of the larger super-eruptions that will undoubtedly occur.[5] [6]

Author: Mt. Rainier, 5/18/1980

The economic contagion observed in recent events such as the 9/11 terrorists attacks, the global financial crises, the Tōhoku earthquake and tsunami, the current Syrian Civil War, and the ongoing Ebola Epidemic in West Africa serve as reasonable real-world benchmarks from which to make projections for much larger events, such as a super-eruption or asteroid. It is not alarmist to state that we are woefully unprepared and need to employ AI to increase the probability for preserving our species and others. The creation and colonization of highly adaptive outposts in space designed with the intent of sustainability would therefore seem prudent policy best served with the assistance of AI.

The role of AI in mitigating the risks associated with super volcanoes include accelerated research, more accurate forecasting, increased modeling for preparation and mitigation of damage, accelerating discovery of systems for surviving super eruptions, and to assist in recovery. These tasks require ultra high scale data analytics to deal with complexities far beyond the ability of humans to process within circumstantially dictated time windows, and are too critical to be dependent upon a few individuals, thus requiring highly adaptive AI across distributed networks involving large numbers of interconnected deep earth and deep sea sensors, linking individuals, teams, and organizations.[7] Machine learning has the ability to accelerate all critical functions, including identifying and analyzing preventative measures with algorithms designed to discover and model scenario consequences.

Asteroids are similar to super volcanoes from an AI perspective, both requiring effective discovery and preparation, which while ongoing and improving could benefit from AI augmentation. Any future preventive endeavors may require cognitive robotics working in extremely hostile environments for humans where real time communications could be restricted. Warnings may be followed swiftly by catastrophic events—hence, preparation is critical, and if viewed as optional, we do so at our own peril.

---

[4] For more common events, see Ultra distant deposits from super eruptions: Examples from Toba, Indonesia and Taupo Volcanic Zone, New Zealand. - N.E. Matthews, V.C. Smith, A Costa, A.J. Durant, D. M. Pyle and N.G.J. Pearce.

[5] USGS Fact Sheet for Alert-Notification System for Volcano Activity: http://pubs.usgs.gov/fs/2006/3139/fs2006-3139.pdf

[6] Paradise is about 45 miles to the north/blast side of St. Helens in full view until ash cloud enveloped Mt. Rainier with large fiery ash and friction lightening. A PBS video on the St. Helens eruption: http://youtu.be/-H_HZVY1tT4

[7] Researchers at Oregon State University claimed the first successful forecast of an undersea volcano in 2011 with the aid of deep sea pressure sensors: http://www.sciencedaily.com/releases/2011/08/110809132234.htm

Pandemics and Healthcare Economics

Among the greatest impending known risks to humans are biological contagions, due to large populations in urban environments that have become reliant on direct public interaction, regional travel by public transportation, and continuous international air travel. Antibiotic-resistant bacteria and viruses in combination with large mobile populations pose a serious short-term threat to humanity and the global economy, particularly from an emergent mutation resistant to pre-existing immunity or drugs.

> *"For example, if a new strain of avian flu were to spread globally through the air travel network that connects the world's major cities, 3 billion people could potentially be exposed to the virus within a short span of time."* – Global Risk Economic Report 2014, World Economic Forum (WEF).[8]

Either general AI or knowledge work augmented with AI could be a decisive factor in preventing and mitigating a future global pandemic, which has the potential capacity to dwarf even a nuclear war in terms of short-term casualties. Estimates vary depending on strain, but a severe pandemic could claim 20-30% of the global population compared to 8-14% in a nuclear war, though aftermath from either could be as horrific.

Productivity in life science research is poor due to inherent physical risks for patients as well as non-physical influences such as cultural, regulatory, and financial that can increase systemic, corporate, and patient risk. Healthcare cost is among the leading global risks due to cost transference to government. The WEF, for example, cites 'Fiscal Crisis in Key Economies' as the leading risk in 2014. Other than poor governance that allows otherwise preventable crises to occur, healthcare costs are arguably the world's most severe curable ongoing crisis, impacting all else, including pandemics.

Priority uses of AI for pandemics and healthcare:[9]

- **Accelerated R&D** throughout life science ecosystem, closing patient/lab gap.
- **Advanced** modeling scenarios for optimizing prevention, care, and costs.
- **Regulatory** process including machine learning and predictive modeling.
- **Precision healthcare** tailored to each person; more direct and automated.[10]
- **Predictive analytics** for patient care, disease prevention, and economics.
- **Cognitive computing** for diagnostics, patient care, and economics.
- **Augmentation** for patients and knowledge workers worldwide.

> *"By contrast to pandemics, artificial intelligence is not an ongoing or imminent global catastrophic risk. However, from a long-term perspective, the development of general artificial intelligence exceeding that of the human brain can be seen as one of the main challenges to the future of humanity (arguably, even the main challenge). At the same time, the successful deployment of superintelligence could obviate many of the other risks facing humanity."* -- Eliezer Yudkowsky in GCR chapter '*Artificial Intelligence as a Positive and Negative Factor in Global Risk*'.

[8] Global Risk Economic Report 2014, World Economic Forum: http://www3.weforum.org/docs/WEF_GlobalRisks_Report_2014.pdf

[9] While pandemics and healthcare obviously deserve and receive much independent scrutiny, understanding the relativity of economics is poor. Unlike physics, functional metric tensors do not yet exist in economics.

[10] New paper in journal *Science*: 'The human splicing code reveals new insights into the genetic determinants of disease' http://www.sciencemag.org/content/early/2014/12/17/science.1254806

Climate Change

Crises often manifest over time as a series of compounding events that either could not be predicted, predictions were inaccurate, or were not acted upon. I've selected an intentionally contrarian case with climate change to provide a glimpse of potentially catastrophic unintended consequences.[11]

> *Anthropogenic climate change has become the poster child of global threats. Global warming commandeers a disproportionate fraction of the attention given to global risks.* - GCR

If earth cooling from human response driven by populism, conflicted science, or any other reason resulted in an over-correction, or a response combined with volcanoes, asteroids, or slight variation of cyclic changes in Earth's orbit to alter ocean and atmospheric patterns, it becomes plausible to forecast a scenario of an accelerated ice age where ice sheets rapidly reach the equator similar to what may have occurred during the Cryogenian period. Given our limited understanding of triggering mechanisms to include in sudden temperature changes, which are believed to have occurred 24 times over the past 100,000 years, it is also plausible that either a natural and/or human response could result in rapid warming. While either scenario of climate extreme may seem ludicrous given current sentiment, the same was true prior to most events. Whether natural, human caused, or some combination as observed with the Fukushima Daiichi nuclear disaster, the pattern of complexity of seemingly small events leading to larger events is more the norm rather than exception, with outcomes opposed to intent not unusual. While reduction of pollutants is prudent for health, climate, and economics, chemically engineered solutions should be resisted until our ability to manage climate change is proven with rigorous real-world models.

9/11

The tragedy of 9/11 provides a critical lesson in seemingly low risk series of events that can easily spiral into catastrophic scale if not planned and managed with precision. Those who decided not to share the Phoenix memo at the FBI with appropriate agencies undoubtedly thought it was a safe decision given information at the time. Who among us would have acted differently if having to make our decision based on the following questions in the summer of 2001?

A. What are the probabilities of an airline being used as a weapon?
B. Even if such an attempt was made, what are the probabilities of it occurring?
C. In the unlikely event such an attempt did occur, what is the probability that one of the towers at the World Trade Center would be targeted and struck?
D. In the very unlikely event one WTC tower were struck, what is the probability that the other tower would be struck similarly before evacuation could take place?
E. In the extremely unlikely scenario that terrorists attacked the second tower with a hijacked commercial airliner within 20 minutes of the first attack, what would be the probability of one of the towers collapsing within an hour?
F. If then this horrific impossible scenario were actually to have occurred, what then would be the possibility of the second tower collapsing a half hour later?

[11] Skeptical views are helpful in reducing the number and scope of unintended consequences. For a rigorous contrarian view on climate change research see 'Fat-Tailed Uncertainty in the Economics of Catastrophic Climate Change', by Martin L. Weitzman: http://scholar.harvard.edu/files/weitzman/files/fattaileduncertaintyeconomics.pdf

Prior to 9/11, such a scenario would seem to be an extremely low probability for anyone but an expert if possible to verify and process all information, but of course the questions describe the actual event, tragically providing one of the most painful and costly lessons in human life and treasure in the history of the U.S., with wide ranging global impact.

- **Direct costs from attacks:** 2,999 lives lost with an additional 1,400 rescue workers having died since, some portion of which were directly caused. Financial costs from 9/11 are estimated to have been approximately $2 trillion to the U.S. alone.[12]
- **Iraq and Afghanistan wars**: Approximately 36,000 people from the US and allied nations died in Iraq and Afghanistan, with nearly 900,000 disability claims approved by the VA alone.[13] The human cost in the two nations were of course much higher, including large numbers of civilians. According to one report, the Iraq War has cost $1.7 trillion with an additional $490 billion in benefits owed to veterans, which could grow to more than $6 trillion over the next four decades counting interest.[14] The Afghanistan war is estimated to have cost $4-6 trillion.
- **Global Financial Crisis:** Many experts believe monetary easing post 9/11 combined with lax regulatory oversight expanded the unsustainable housing bubble and triggered the financial collapse in 2008, which revealed systemic risk accumulating over decades from poor regulatory oversight, governance, and policy.[15] The total cost to date from the financial crisis is well over $10 trillion and continues to rise from a series of events reverberating globally, caused by realized moral hazard, broad global reactions, and significant continuing aftershocks from attempts to recover.

The reporting process for preventing terrorist attacks has been reformed and improved, though undoubtedly still challenged due to the inherently conflicted dynamics between interests, exponential growth of data, and limited ability for individuals to process. If the FBI decision makers were empowered with the knowledge of the author of the Phoenix memo (special agent Kenneth Williams), and free of concerns over interagency relations, politics, and other influences, 9/11 would presumably have been prevented or mitigated.

An Artificially Intelligent Special Agent

What if the process wasn't dependent upon conflicted humans? An artificial agent can be programmed to be free of such influences, and can access all relevant information on the topic from available sources far beyond the ability of humans to process. When empowered with machine learning and predictive analytics, redundant precautions, and integrated with the augmented workflow of human experts, the probability of preventing or mitigating human caused events is high, whether in preventing terrorist plots, high cost projects with deep sea oil rigs, trading exotic financial products, nuclear power plant design, response to climate change, or any other such high-risk endeavor.

---

[12] Institute for the Analysis of Global Security http://www.iags.org/costof911.html

[13] US and Coalition Casualties in Iraq and Afghanistan by Catherine Lutz, Watson Institute for International Studies: http://costsofwar.org/sites/default/files/articles/8/attachments/USandCoalition.pdf

[14] Daniel Trotta at Reuters, March 14th, 2013 http://www.reuters.com/article/2013/03/14/us-iraq-war-anniversary-idUSBRE92D0PG20130314

[15] In his book *The Age of Turbulence: Adventures in a New World*, Alan Greenspan states that monetary policy in the years following 9/11 "saved the economy", while others point to Fed policy during this period as contributing to the global financial crisis beginning in 2008, which continues to reverberate today. Government guarantees, high-risk products in banking, companies 'too big to fail', corrupted political process, increased public debt, lack of effective governance, and many other factors also contributed to systemic risk and the magnitude of the crisis.

Governance of Human Engineered Systems

Long considered one of the primary technical risks, as well as key to overcoming many of our most serious challenges, the ability to program matter at the atomic scale is beginning to emerge.[16] Synthetic biology is another area of modern science that poses great opportunity for improving life while introducing new risks that must be governed wisely with appropriate tools.[17]

While important private research cannot be revealed, published work is underway beyond the typical reporting linked to funding announcements and ad budgets. One example is the *Synthetic Evolving Life Form* (SELF) project at Raytheon led by Dr. Jim Crowder, which mimics the human nervous system within an *Artificial Cognitive Neural Framework* (ACNF).[18] In their book *Artificial Cognitive Architectures,* Dr. Crowder and his colleagues describe an unplanned Internet discussion between an artificial agent and human researcher that called for embedded governance built into the design.

A fascinating project was recently presented by Joshua M. Epstein, Ph.D. at the Santa Fe Institute on research covered in his book *Agent_Zero: Toward Neurocognitive Foundations for Generative Social Science*.[19] In this novel approach to agent based social behavior, Epstein has intentionally removed memory and pre-conditioned bias from Agent Zero to simulate what often appears to be irrational behavior resulting from social influences, reasoning, and emotions. Agent Zero provides a good foundation for considering additional governance methods for artificial agents interacting in groups, with potential application in AI systems designed to regulate systemic risk in networked environments.

Closing Thoughts

Given the rapidly developing AI-assisted world combined with recent governance failures in financial regulation, national security, and healthcare, the challenge for senior decision makers and AI systems engineers is to perform at much higher levels than the past, for if we continue on this trajectory it is conceivable that the human experience could confirm the Fermi Paradox within decades. I therefore view AI in part as a race against time.

While the time-scale calculus for many organizations supports accelerated transformation with the assistance of AI, it is wise to proceed cautiously with awareness that AI is a transdisciplinary field which favors patience and maturity in the architects and engineers, typically requiring several tens of thousands of hours of deep immersion in each of the applicable focus areas. It is also important to realize that the fundamental physics involved with AI favors if not requires the governance structure to be embedded within the engineered data, including business process, rules, operational needs, and security parameters. When conducted in a prudential manner, AI systems engineering is a carefully planned and executed process with multiple built-in redundancies.

---

[16] A team of researchers published research in Nature Chemistry in May of 2013 on using a Monte Carlo simulation to demonstrate engineering at the nanoscale: http://www.nature.com/nchem/journal/v5/n6/full/nchem.1651.html

[17] A recent paper from MIT researchers 'Tunable Amplifying Buffer Circuit in *E. coli'* is an example of emerging transdisciplinary science with mechanical engineering applied to bacteria: http://web.mit.edu/ddv/www/papers/KayzadACS14.pdf

[18] My review of Dr. Crowder's recent book *Artificial Cognitive Architectures* can be viewed here: http://kyield.wordpress.com/2013/12/17/book-review-artificial-cognitive-architectures/

[19] *Agent_Zero: Toward Neurocognitive Foundations for Generative Social Science, by* Joshua M. Epstein: http://press.princeton.edu/titles/10169.html The seminar slides (require Silverlight plugin): http://media.cph.ohio-state.edu/mediasite/Play/665c495b7515413693f52e7ef9eb4c661d

Mark Montgomery is founder and CEO of http://www.kyield.com, which offers technology and services centered on Montgomery's AI systems invention.

(Robert Neilson, Ph.D., and Garrett Lindemann, Ph.D., contributed to this article)