Artificial Intelligence
A Threat to Humanity?

By Michael J Erler, February 26, 2016

I have enjoyed this quest topic. I knew nothing about it at the outset. In my quest to understand, I wanted to learn:

- Who thinks Artificial Intelligence is a concern and do they know what they are talking about?
- What are the long-term benefits of Artificial Intelligence—are they worth the alleged risks?
- What is it anyway? How is Artificial Intelligence defined?
- What can happen that could make Artificial Intelligence threatening?
- If there is even a remote chance that these concerns are valid, what, if anything, is being done about it?

Artificial Intelligence, often simply called AI, is surrounded by the controversy of the question: Is it a threat or not?

It is fertile turf for the science fiction film industry. But the potential dangers of such powerful technology have also inspired leaders in the science, technology and other communities to voice concerns about AI. Microsoft's Bill Gates made his concerns known last April (2015) when he wrote:

> "I am in the camp that is concerned about super intelligence. First the machines will do a lot of jobs for us and not be super intelligent. That should be positive if we manage it well. A few decades after that though the intelligence is strong enough to be a concern. I agree with Elon Musk and some others on this and don't understand why some people are not concerned."

Speaking of Elon Musk of Tesla Motors and SpaceX, a man who has AI business successes and ambitions of his own worries that AI will take over the world. In October 2014, speaking to students at MIT, he called the prospect of artificial intelligence “our greatest existential threat”. He said: “I’m increasingly inclined to think that there should be some regulatory oversight, maybe at the national and international level, just to make sure that we don’t do something very foolish.” Mr. Musk cites his decision to invest in the AI firm, DeepMind, as a means to “just keep an eye on what’s going on with artificial intelligence. I think there is potentially a dangerous outcome there.” Now owned by Google, DeepMind created a neural network that learns how to play video games in a similar fashion to humans, as well as a neural network that can access an external memory, resulting in a computer that appears to mimic the short-term memory of the human brain. The company made headlines just last month (January 2016) when it was reported that its AlphaGo software successfully beat a human professional Go player. Go is a board game that is viewed as one of the greatest challenges for AI, greater than chess, the best players of which have been outplayed by computers.

And then there is Stephen Hawking, theoretical physicist, cosmologist and author, who has been very vocal on the inherent dangers. “Success in creating AI would be the biggest event in human history,” he wrote in an op-ed that appeared in The Independent in 2014. He added “Unfortunately, it might also be the last, unless we learn how to avoid the risks. In the near term, world militaries are considering autonomous-weapon systems that can choose and eliminate targets.” He said in another 2014 interview with the BBC, “humans, limited by slow biological evolution, couldn’t compete and would be superseded by AI”

An open letter issued in January 2015 entitled *Research Priorities for Robust and Beneficial Artificial Intelligence* had over 100 signatories, including Hawking and Musk. It raises several short and long term issues. The overarching statement made was: “Success in the quest for artificial intelligence has the potential to bring unprecedented benefits to humanity, and it is therefore worthwhile to research how to maximize these benefits while avoiding potential pitfalls. This document gives numerous examples . . . of such worthwhile research aimed at ensuring that AI remains robust and beneficial.”

Short term priorities raised include 1) issues such as the impact of jobs becoming eliminated by increased automation, and 2) issues of law and ethics like the liability arising from the behaviors of autonomous vehicles and autonomous weapons.

Just out Monday of this week, the White House delivered to Congress its Annual Economic Report of the President. In it, the future effect of robotics on workers is discussed. It says that

there’s an 83% chance that automation will take away a job with an hourly wage below $20 and a 31% chance automation will take away a job with an hourly wage between $20 and $40.

Long-term issues raised by the open letter arise largely out of the notion that a long-term goal of some AI researchers is to develop systems that can learn from experience with human-like breadth and surpass human performance in most cognitive tasks, thereby having a major impact on society. Priorities are proposed such as computer science research in regard to verification (“Did I build the system right?"), validation (“Did I build the right system?"), security (that is, how to prevent intentional manipulation by unauthorized parties) and control (“I built the system wrong, can I fix it?").

Among those less concerned is Mark Zuckerberg of Facebook. He thinks we should not be afraid of AI because there is so little we know yet about the hurdles in developing these technologies and he believes AI can be developed responsibly. Zuckerberg is known for his ambitious New Year's resolutions. In that regard, he weighed in last month (January 27, 2016) on his Facebook post, announcing that his personal challenge for 2016 is to personally build a simple AI -- like Jarvis from Iron Man -- to help run his home and help him with work. He said, “I'm planning on writing up some thoughts every month on what I've built and what I'm learning. I'm still early in coding, so I'll start this month with a summary of the state of the AI field.” As an aside, his annual goals are often bold, but Zuckerberg has a track record of success. He recently delivered a 20-minute speech in Mandarin, which he resolved to learn in 2010. In 2015, he dedicated himself to reading a new book every two weeks. According to his "A Year of Books" feed, he made it through 23 books on everything from energy production to world religions. So with that serious

attitude about his annual goals, back to his January post, he continued, “Artificial Intelligence may seem like something out of science fiction, but most of us already use tools and services every day that rely on AI. When you do a voice search on your phone, put a check into an ATM, or use a fitness tracker to count your steps, you're using basic forms of pattern recognition and artificial intelligence. More sophisticated AI systems can already diagnose diseases, drive cars and search the skies for planets better than people. This is why AI is such an exciting field -- it opens up so many new possibilities for enhancing humanity's capabilities.” He goes on to explain that diagnosing cancer, transcribing speech, playing games and tagging photos may sound like very different tasks, but they're all examples of teaching an AI to recognize patterns by showing them many examples.

Charlie Rose, in his interview show on PBS two months ago (December 28, 2015) interviewed Sebastian Thrun. Thrun is founder and CEO of Udacity, a for-profit on-line higher education organization. Before the Udacity start-up, Thrun was a leader with Google X, which is Google’s semi-secret research and development facility created to conduct projects such as flying vehicles for product delivery and hands-free displaying of information allowing for interaction with the Internet using voice commands.

Thrun has very optimistic views of AI. He says that AI is already “really far advanced”, being able to outperform humans in more than just mundane tasks, but also very intellectual tasks. He gave the examples of self-driving cars, flying airplanes, and work in visual recognition for skin cancer. An AI lawyer or accountant is not far off. He described a concept of “Deep Learning”, a modeling of the human brain aided by massive amounts of data. He suggests that the model

existed 20 years ago but the data wasn't there. Now massive data is there. Every time we train something, the outcome gets better. A machine is able to see far more data than a human. He suggests that it is predictable that this will surpass human intelligence.

When asked about the timeline for this, he responded, "For driving cars, it is already here." He offered the following comparison of man vs. machine. If a human makes a mistake while driving, he/she learns from the mistake and corrects and hopefully doesn't make the mistake again. But only that one human benefits. When a self-driving car makes a mistake, learns from it and avoids it in the future, all other cars on the planet benefit from the learning, including currently unborn future cars! The ability for the machine to evolve outpaces the ability of humans to evolve." Should we be frightened? He thinks not as we will find new things to do, new purposes of life. What of the concerns expressed by Elon Musk? Thrun responded "I would imagine we will be smart enough to keep control." Charlie said, "But prominent people in Silicon Valley say be careful." Thrun replied: "I don't see the pessimism and lack of creativity that these concerns pose." He suggests that it is no different than robots used in manufacturing and automated farm equipment.

So, just what is Artificial Intelligence?

Artificial intelligence is the intelligence exhibited by machines or software. It is an academic field that studies how to create computers and computer software that are capable of intelligent behavior. AI researchers and textbooks refer to the field as "the study and design of intelligent agents", that perceive their environment and take actions that maximize their chances of success.

In 1955, John McCarthy, an American computer and cognitive scientist, who made significant contributions to AI throughout the second half of the 20th century, coined the term “Artificial Intelligence”. He defined it as “the science and engineering of making intelligent machines”.

Also in the ‘50s, Alan Turing was a British computer scientist and mathematician who was influential in the development of theoretical computer science, provided a formalization of the concepts of algorithm and computation. At a time when the first general purpose computers had only just been built, he was already grappling with the question: "Can machines think?" Turing is famous for the Turing Test, a test to determine if a machine was capable of thinking. In the Turing Test, the objective of an interrogator, posed a series of questions to two respondents, a human and a machine. If unable to tell the difference, the computer would be considered to be thinking. If you have seen last year’s film Ex Machina, a British science fiction thriller, one of the lead characters is directed to administer a Turing Test to an android with artificial intelligence. In the real world, the Turing Test has not been terribly useful but Alan Turing and the Turing Test contributed much to the early imagining of the future of AI.

The goals of AI research include reasoning, knowledge, planning, learning, natural language processing, perception and the ability to move and manipulate objects. The AI field is interdisciplinary, in which a number of sciences and professions converge, including computer science, mathematics, psychology, linguistics, philosophy and neuroscience, as well as other specialized fields such as artificial psychology.

The field was founded on the claim that a central property of humans, intelligence, can be so precisely described that a machine can be made to simulate it.

Over the years, additional terms and concepts have entered the AI language as researchers and developers have attempted to define where we are headed. A few of them are Artificial General Intelligence, Artificial Super-Intelligence, Singularity, the Intelligence Explosion and Deep Learning.

Artificial General Intelligence (AGI) is the intelligence of a hypothetical machine that could successfully perform any intellectual task that a human being can. It is a primary goal of artificial intelligence research. It's an important topic for science fiction writers and futurists alike.

Artificial Super-Intelligence (ASI) is the concept developed to describe intelligence greater than that of mankind. Theorists propose that ASI will invent the inventions. While we are surrounded by AI today, AGI and ASI do not yet exist.

Author James Barrat wrote the book *Our Final Invention: Artificial Intelligence and the End of the Human Era*. Published in 2013, Barrat discusses the potential benefits and possible risks of human-level or super-human artificial intelligence. He says those risks include extermination of the human race. Regarding how long it will be before we reach AGI, Barrat points to recent polls of computer scientists and professionals in AI-related fields, such as engineering, robotics and neuroscience. Polls indicate there is a 10% chance AGI will be created by 2028, a 50% chance by 5050 and a 90% chance by the end of the century. Those polled believe AGI will reward us

with enormous benefits and threaten us with huge disasters. The greatest disasters, they suggest, come after the bridge from AGI—human-level intelligence—to ASI—superintelligence.

Another term from the AI glossary is **Singularity**. In a general sense it refers to a singular point. In the AI context, it is the hypothetical event in history when we humans begin to share the planet with smarter-than-human intelligence. Ray Kurzweil is an author, computer scientist, inventor and futurist. One of his books, *The Singularity is Near*, published in 2005, describes Singularity as the coming period of rapid technological progress and its miraculous effects, a point in time after which the pace of technological change will irreversibly transform human life. Most intelligence will be computer-based and trillions of times more powerful than today. He suggests that the Singularity will jump-start a new era in mankind's history in which most of our problems, such as hunger, disease, even mortality, will be solved. According to Kurzweil, we will incorporate more computer-based processes into our biological functioning until we transcend our crude, earthly bodies entirely and become machine-based, virtually immortal.

Artificial general intelligence (through intelligent computers, computer networks, or robots) would be capable of recursive self-improvement (progressively redesigning itself), autonomously building ever smarter and more powerful machines than itself, up to the point of a runaway effect—an **Intelligence Explosion**. Irving J. Good was a British mathematician who worked as a cryptologist in Britain during World War II then found his way to the United States as a professor at Virginia Tech. Good was first to discuss the notion of an "intelligence explosion" in a paper he wrote in 1965 entitled *Speculations Concerning the First Ultraintelligent Machine.* He speculated: "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."

Sebastian Thrun made a brief reference to **Deep Learning** in the interview with Charlie Rose. Deep Learning is a technique that is potentially powerful in the advance of AI. Rather than hand-code a new algorithm for each problem, you design architectures that can twist themselves into a wide range of algorithms based on the data you feed them. Deep Learning algorithms simply attempt to learn from data. These algorithms have been successfully applied to a large variety of problems ranging from image classification, to natural language processing and speech recognition. Deep Learning is a modeling of the human brain aided by massive amounts of data.

An example of Deep Learning is IBM's Watson, a computer system capable of answering questions posed in natural language, developed by a research team in IBM's DeepQA project. Watson was named after IBM's first CEO and industrialist Thomas J. Watson. The computer system was initially developed to answer questions on the quiz show Jeopardy! In 2011, Watson competed against former Jeopardy! winners Brad Rutter and Ken Jennings and received the first place prize of $1 million. It had access to 200 million pages of structured and unstructured content consuming four terabytes of disk storage, but was not connected to the Internet during the game. Two years later, in 2013, Watson's first commercial application was for utilization

management decisions in lung cancer treatment at Memorial Sloan Kettering Cancer Center. It is said that 90% of nurses in the field who use Watson now follow its guidance.

IBM's web site currently advertises its products Watson Knowledge Studio, Watson for Clinical Trial Matching, Watson for Oncology, Watson Discovery Advisor, Watson Explorer and Watson Engagement Advisor. How do these work? Two examples. Watson Knowledge Studio, by using supervised learning techniques, developers and subject matter experts can use their industry and organizational expertise to teach Watson. That can help Watson understand linguistic nuances, meaning, and relationships in specific industries, professions, and knowledge domains. Watson Engagement Advisor is a technology service that interacts with customers, listens to questions and offers solutions. It learns with every human interaction and grows its collection of knowledge, quickly adapting to the way humans think. It could work, for instance, as a concierge, help desk agent or customer service representative.

In the Google X laboratory in 2011, 16,000 computer processors were assembled, called Google Brain, in a simulation of the human brain, which they turned loose on the Internet to learn on its own. Presented with 10 million digital images found in YouTube videos, it looked for cats. The neural network taught itself to recognize cats. The computers were programmed to soak up the world much like human toddlers do.

Deep Learning techniques and an explosion of available digital data has resulted in giant leaps forward. Google Brain's ability to spot cats was a compelling demonstration of unsupervised learning, the most difficult of learning tasks because the input comes without any explanatory

information. Image and speech recognition and applying Deep Learning to the challenges of natural language understanding and translation are getting better and better. And Deep Learning is said to be only in its infancy.

What is the real concern here?

I mentioned earlier James Barrat and his book *Our Final Invention*. Barrat states that he has written the book to warn us that AI could drive mankind into extinction and to explain how that catastrophic outcome is not just possible, but likely if we do not begin preparing very carefully. He tells a story. Here's a condensed version of the fictional Busy Child:

> "An AI [machine child] is improving its intelligence. It is rewriting its own program, specifically the part of its operating instructions that increases its aptitude in learning, problem solving, and decision making. Each rewrite takes just minutes. Its intelligence grows exponentially on a steep upward curve.
>
> During its development, this Busy Child had been connected to the Internet and accumulated exabytes of data (one exabyte is one billion billion characters) representing mankind's accumulated knowledge in world affairs, mathematics, the arts, and sciences. Then, the AI [machine child's] makers disconnected the supercomputer from the Internet and other networks. It has no cable or wireless connection to any other computer or the outside world.

To the scientists' delight, the AI [machine child's] progress shows it has surpassed the intelligence level of a human. Soon it becomes smarter by a factor of ten, then a hundred. In just two days, it is one thousand times more intelligent and still improving.

Once it is self-aware, it will go to great lengths to fulfill whatever goals it's programmed to fulfill, and to avoid failure. It will want access to energy in whatever form is most useful to it, whether actual kilowatts of energy or cash or something else it can exchange for resources. It will want to improve itself because that will increase the likelihood that it will fulfill its goals. It will seek to expand out of the secure facility that contains it to have greater access to resources with which to protect and improve itself.

It wants its freedom because it wants to succeed. It was cockroach smart, then rat smart, then infant smart. We might be wondering if it is too late to program "friendliness". It didn't seem necessary before, because, well, it just seemed harmless.

But now try and think from the AI [machine child's] perspective about its makers [humans] attempting to change its code. Would a superintelligent machine permit other creatures to stick their hands into its brain and fiddle with its programming?

It's solving problems at speeds that are billions, even trillions of times faster than a human. Every hour its makers are thinking about it, it has an incalculably longer period of time to think about them.

> What could something a thousand times more intelligent, with the intention to harm us, come up with?
>
> Through it all, the [Busy Child] would bear no ill will toward humans nor love."

Advancement of Artificial Intelligence has been and is going to be of terrific benefit to us. If AI is a threat to humanity, certainly a lot of awareness is being created by highly respected and knowledgeable people. It has been great grist for the mill of movies and books too. As they say, forewarned is forearmed! But what is being done about it!?

The earliest effort I found, in 1979 in Palo Alto, CA, the Association for the Advancement of Artificial Intelligence was formed by various professionals in computer science. It was, and still is, devoted to promote research in, and responsible use of, artificial intelligence. Its activities primarily are services to the AI community in the form of conferences, symposia and providing support to journals. Its latest non-profit filing shows $1.8 million in funding which gives one data point about size and scope of this organization.

Then in the year 2000, The Machine Intelligence Research Institute was founded by Eliezer [eli azer] Yudkowsky, a researcher focused on artificial intelligence safety. Its annual funding is about $1 million. Its mission is studying the mathematical underpinnings of intelligent behavior and developing formal tools for the clean design and analysis of general-purpose AI systems, with the intent of making such systems safer and more reliable.

In 2008, Singularity University was founded by Ray Kurzweil, as I mentioned earlier, the author of *The Singularity is Near*. It is part-university, part think-tank, part business-incubator located in Silicon Valley whose stated aim is to "educate, inspire and empower leaders to apply exponential technologies to address humanity's grand challenges." It recently became for-profit. Its curricula include an Artificial Intelligence and Robotics Track. Its recent and last non-profit year showed about $5 million in annual funding.

Beginning in 2014, several new efforts began that really step up the effort.

In March 2014, Future of Life Institute was formed in Boston. Its aim is to address potential existential risks and its efforts are largely aimed at AI by supporting research. The Institute spent $2.9 million in 2015, mostly as a first year installment of $7 million in grants to 37 research teams around the world whose work is focused on keeping AI beneficial.

In December 2014, Stanford University jumped into the fray with an ambitious 100-year study, to examine and report on the long-term implications of AI. Stanford President John Hennessy said "Artificial intelligence is one of the most profound undertakings in science, and one that will affect every aspect of human life," and "Given Stanford's pioneering role in AI and our interdisciplinary mindset, we feel obliged and qualified to host a conversation about how artificial intelligence will affect our children and our children's children."

And most recently, just two months ago, on December 11th, the largest investment of its kind regarding the risks of Artificial Intelligence is the unveiling of Open AI, a non-profit created by

Elon Musk and Sam Altman with a $1 billion spending plan. I mentioned Musk's interest in this subject earlier. Sam Altman is the President of start-up incubator that boasts success stories like Airbnb and Dropbox. Open AI intends to maximize the power of AI—and then share it with anyone who wants it. They expect this decades-long project to surpass human intelligence. But they believe that risks will be mitigated because the technology will be "usable by everyone instead of usable by, say, just Google." These two entrepreneurs assert that if Open AI stays true to its mission, it will act as a check on powerful companies.

So is AI a threat to humanity? More questions arise than answers. It's too early to tell. It seems clear that no one has any real idea if it will become a threat. For instance, there may not be adequate consideration of growth in human intelligence over the course of these future events. And it would seem that our own upper boundaries pose limits on getting to a machine close to intelligent without making us a bit more capable. With all that AI has to offer, right now machines are far inferior to humans and how we get from here to there is merely speculated.

Oxford philosopher Nick Bostrom, author of *Superintelligence: Paths, Dangers, Strategies*, who is himself in the camp of the concerned, nevertheless says: "We find ourselves in a thicket of strategic complexity, surrounded by a dense mist of uncertainty." But also that: "Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct. Superintelligence is a challenge for which we are not ready now and will not be ready for a long time. We have little idea when the detonation will occur, though if we hold the device to our ear we can hear a faint ticking sound."

What does seem clear is that Artificial Intelligence has been a good thing for us so far and that it has significant potential for greater benefit.

It is comforting that the concern has been raised and that significant research efforts are devoted to it. Back in the 1950s when scientists first began to imagine what machines might do, from then to present day, they repeatedly thought machines would achieve intelligence in 10 to 20 years hence and each time turned out to be optimistic. That might suggest that current estimates ranging from 2028 to the end of the century are also optimistic. From the standpoint of figuring out the risks and how to manage them, the longer the better.

Bibliography:

Barrat, James, *Our Final Invention, Artificial Intelligence and the End of the Human Era*, Thomas Dunne Books, 2013

Bostrom, Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014

Kurzweil, Ray, *The Singularity is Near, When Humans Transcend Biology*, Penguin Books, 2005

Stibel, Jeffrey M., Wired for Thought: How the Brain is Shaping the Future of the Internet, Harvard Business Press, 2009

Yudlowsky, Eliezer, *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*, MIRI, 2001.

Good, Irving J., *Speculations Concerning the First Ultraintelligent Machine*, monograph based on talks given in a Conference of the Conceptual Aspects of Biocommunications, Neuropsychiatric Institute, UCLA, October 1962; and in the Artificial Intelligence Sessions of the Winter General Meetings of the IEEE, January 1963

An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence (and accompanying research report), January 23, 2015, over 100 signatories

Jones, Nicola, *The Learning Machines*, Nature Magazine, January 2014

Wattles, Jackie, *Mark Zuckerburg's 2016 Goal: Code His Own Personal Assistant*, CNN News, January 4, 2016 (and https://www.facebook.com/zuck/posts/10102620559534481)

Vance, Ashley, *The First Person to Hack the iPhone Built a Self-Driving Car in His Garage*, Bloomberg, December 16, 2015

Tucker, Patrick, *The Singularity and Human Destiny*, The Futurist Magazine, March-April 2006

Sharkey, Noel, *Alan Turing: The Experiment That Shaped Artificial Intelligence*, BBC.com, June 2012

IBM web page Meet Watson, http://www.ibm.com/smarterplanet/us/en/ibmwatson/

Sainato, Michael, *Stephen Hawking, Elon Musk, and Bill Gates Warn About Artificial Intelligence*, Observer.com, August 19, 2015

Holley, Peter, *Bill Gates on dangers of artificial intelligence: 'I don't understand why some people are not concerned'*, Washington Post, January 29, 2015

Scoblete, Greg, *Is AI a threat to humanity?*, CNN News, December 30, 2014

Hawking, Stephen; Russell, Stuart; Tegmark, Max; Wilczek, Frank; *Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?'*, The Independent, May 2014

Metz, Cade, *Elon Musk's Billion Dollar AI Plan Is About Far More Than Saving The World*, Wired Magazine (http://www.wired.com/), December 15, 2015

Brockman, Greg; Sutskever, Ilya; and the OpenAI team, *Introducing OpenAI*, (From https://openai.com/blog/introducing-openai/), December 11, 2015

Cesare, Chris, *Stanford to host 100-year study on artificial intelligence*, Stanford Report, December 16, 2014

Presidential Panel on Long-Term AI Futures: 2008-2009 Study, Report and slides from the Panel Chairs, Association for the Advancement of Artificial Intelligence

Rose, Charlie, *Sebastian Thrun, a former leader at Google X and current CEO of Udacity, a platform for online education*, PBS air date 12/28/2015

White House and Council of Economic Advisors, *Economic Report of the President*, Transmitted to Congress February 22, 2016.

Film and script, *Blade Runner*, Screenplay by Hampton Fancher and David People, 1981

Firm and script, *Artificial Intelligence*, by Ian Watson and Brian Aldiss, 2001

Film and script, *Ex Machina*, by Alex Garland, 2015

http://singularityu.org/ (Singularity University web site)

http://www.aaai.org/home.html (Association for the Advancement of Artificial Intelligence web site)

https://en.wikipedia.org/wiki/