Artificial Superintelligence: Extinction or Nirvana?

Jens Pohl, PhD
Professor of Architecture (Emeritus)
California Polytechnic State University (Cal Poly)
Senior Director, Adaptive Systems
Tapestry Solutions (a Boeing Company)
San Luis Obispo, California, USA

Abstract

The purpose of this paper is to explore the subject of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI) and identify in summary form the principal drivers and current status of Artificial Narrow Intelligence (ANI), the impact that intelligent software in combination with the Internet serving as a Global Knowledgebase is already having on human capabilities, the likely path to ASI, the possible threats posed by ASI and proposed measures to curtail those threats, and finally the potential benefits to the human species if AGI and ASI remain under human control.

Keywords

artificial general intelligence, AGI, artificial intelligence, AI, artificial narrow intelligence, ANI, artificial superintelligence, ASI, global knowledgebase, intelligence explosion, singularity

Introduction and Definitions

While the initial high expectations for artificial intelligence capabilities were not realized in the proposed early timeframe, there is now almost general agreement that such capabilities are achievable and that they will have a profound impact on the human species. Some of these impacts are already apparent and are changing the way we acquire knowledge, how we work, and how we communicate. We are rapidly becoming dependent on computers to outsource our memory and on the Internet to augment our knowledge.

The computer is becoming an intelligent machine and there are indications that its level of intelligence may eventually surpass that of its human creator. We may well be heading toward an artificial *Intelligence Explosion*. If such a scenario comes true then Artificial Intelligence (AI) will transition in rapid succession to what is now commonly referred to as *Artificial General Intelligence (AGI)* and then exponentially to *Artificial Superintelligence (ASI)*. AGI will have equal or superior intelligence to the human in most domains and ASI is expected to be many orders of magnitude more intelligent than the human.

Within this context there has emerged a relatively small but growing group of prominent computer and cognitive scientists, philosophers, sociologists, and psychologists who believe that this unavoidable transition from AI to AGI to ASI is fraught with danger and may in fact lead to the eventual demise of the human species. They readily acknowledge that throughout human history major technical accomplishments such as the printing press and the mechanical machine were initially met with major concerns and that eventually those concerns turned out to be unwarranted. However, they also point out that none of these innovations dealt with the main reason for mankind's dominance, namely intelligence. On the other hand, it must also be recognized that there is a distinct pattern on planet Earth that species eventually become extinct. Urban (2015b) refers to this as the *life balance beam* and suggests that historically 99.9% of all species have fallen off the balance beam. It is the fear that ASI might become the force that will push the human species off the *life balance beam* that fuels the debate between those that consider ASI as a serious threat and those that consider ASI as the salvation of the human race that may eventually result in immortality.

Although the emerging subject of AGI and ASI is advancing rapidly its discussion is still largely confined to a relatively small special interest community. It is therefore necessary to define some of the descriptive terms that will be used throughout the paper.

> ***Artificial Narrow Intelligence (ANI):*** - Also referred to as *weak AI* or simply *AI*. This is essentially the state of computer software today. Computer programs are designed and mostly developed by human programmers with the process from input to output open to inspection. Programs typically specialize in one area and can be mathematically proven to be *safe* or *friendly*. Even in model-driven architectures where the source code production is partially automated the resulting code is designed by humans and the code execution results are largely predictable.
>
> ***Artificial General Intelligence (AGI)*** **–** The level of artificial intelligence that has been reached when a computer has equaled the intelligence level of a human in virtually all areas including reasoning, planning, solving problems, thinking abstractly, comprehending complex ideas, learning quickly, and learning from experience.
>
> ***Artificial Superintelligence (ASI)*** **–** The level of artificial intelligence that has been reached when a computer has exponentially surpassed the intelligence level of a human by several orders of magnitude.
>
> ***Intelligence Explosion*** – A point in time when a computer is capable of recursive self-improvement and self-awareness. First defined by I. J. Good (1965) as the ability of an *ultra-intelligent* machine to design even better machines leading to *ASI*.
>
> ***Singularity*** – A term that has been used loosely to describe different aspects of a man-machine relationship that is expected to come into existence when collective human and collective computer intelligence are approximately on par. Therefore Singularity is likely to be reached coincidentally with the *Intelligence Explosion* and *AGI*. It is speculated that in a post-Singularity world there will be little, if any, distinction between physical and virtual reality.

Underlying the terms *Intelligence Explosion*, *AGI*, *Singularity*, and *ASI* is the notion that human-created technology is subject to the Law of Accelerating Returns (LOAR). Kurzweil (2005) has treated this notion extensively, providing convincing evidence that technology is evolving exponentially, and projecting in some detail the potential impact of such an exponential evolution rate on a largely unprepared human species.

Evidence of exponential growth can be found throughout human history. For example, we can arrive at an approximate growth rate of the world economy by extrapolating subsistence-level income in combination with population growth from pre-historic times to the present day (Bostrom 2005, 1-2). While during the earliest stages in the evolution of Homo sapiens it took hundreds of thousands of years to reach a bare subsistence productivity level, following the Agricultural Revolution the same increase in productivity was achieved in less than two centuries and after the Industrial Revolution in no more than a few years. It is therefore not unreasonable to project that with another transition such as the Information Age the world economy could double in size on a monthly or even a weekly basis.

The concept of exponential growth is very difficult for us humans to come to terms with because we are situated in an environment that appears to be governed by linear growth. Both distance and time, which play a fundamental role in virtually all of our daily activities, are perceived by us to be subject to linear rules. The notion of time governed by exponential rules is inconceivable to us. For example, it would mean that on an exponential time scale we would have aged by almost 1,000 years in a single 24-hour day[1].

The Uniqueness of Human Intelligence

As a starting point for a discussion of artificial intelligence it is appropriate to briefly examine some of the principal characteristics of human intelligence. As an organism we humans have five basic senses, namely: sight; hearing; taste; smell; and, touch. At least some minimal combination of these senses allows us to interact with the physical environment and each other. While there are other human senses such as temperature, pain, balance, and kinesthetic sense (i.e., movement of muscles and joints), these are more closely related to survival than intelligence.

The essential components of human intelligence are knowledge and experience, reasoning, learning ability, creativity, intuition, and emotions. Knowledge is acquired through the analysis of information and experience is gained through the accumulation of knowledge over time. Although knowledge is not the same as intelligence, it is a powerful amplifier of intelligence. For example the Google search engine has multiplied worker productivity particularly in occupations that require research and writing (Barrat 2013, 190). Through logical reasoning we convert information into knowledge and solve problems.

Our ability to learn allows us to gain new knowledge and revise existing knowledge based on interactions with the physical environment and fellow human beings. In other words, we learn through the analysis of information, the direct acquisition of knowledge from various external sources, and our experience. While we continue to learn throughout our lifetime, the first quarter of our lifespan is typically subject to a formal process in which learning is accelerated through the selection of subject matter and tutelage. This suggests that human learning is a relatively slow process that is in need of assistance.

Closely associated with learning is the ability to analyze information and solve problems on a thought-based level, commonly referred to as abstract reasoning. This capability also leads to the ability to create new knowledge. Creativity is the most complex form of human intelligence and therefore also its rarest characteristic that is associated with observation, motivation, reasoning, intuition, and experimentation. The most mysterious of these qualities is intuition, which allows humans to reach conclusions without any conscious reasoning. However, these conclusions may be false due to misunderstandings or biased due to influences such as status quo, wishful thinking, or judging new circumstances based on past conditions (Bonabeau 2003).

Finally, all humans are subject to instinctive emotional responses that are intertwined with the mood, temperament, personality, and motivation of the individual. Emotions are a complex state

[1] On a linear scale time advances by exactly one hour every hour. However, on an exponential scale time would advance at a rate that doubles each hour. Therefore, in the second linear hour time would advance by two hours and in the third linear hour by four hours, and so on. In the 24th linear hour time would advance by 8,388,608 hours or 349,525 days or 957 years.

of feelings that influence human behavior. They may positively provide motivation for creativity or negatively subvert reasoning along an illogical path and produce false conclusions or intuitions.

To what extent would or could these characteristics of human intelligence be achieved through machine intelligence? Logical reasoning in combination with unlimited access to information and immense computational speed may be sufficient for a machine to acquire knowledge, gain experience, learn, and create new knowledge. However, the human brain has other sophisticated cognitive capabilities such as complex linguistic representation, long term planning, and abstract reasoning that cannot be reproduced through computational power and speed alone. This is what Urban (2015b) refers to as *intelligence quality*. He gives the example of a chimpanzee that can recognize a building but cannot understand that anyone or anything can construct a building. To the chimpanzee the building is part of the natural environment. Urban then argues that this difference in intelligence quality between the human and the chimpanzee is very small compared with the difference in intelligence quality between ANI and AGI.

Both intuition and emotions might be handicaps rather than desirable capabilities for ANI to transition to AGI. However, it may be argued that for true AGI to emerge there must be an emotional component. Certainly in human decision making emotion is often stronger than logic. This raises the question to what extent will or should AGI emulate human intelligence? This question leads to two distinct considerations, namely: would an emotional component perhaps present the best approach to providing a degree of human influence on defining the goals and objectives of ASI; and, would the absence of an emotional component be an advantage by ensuring that ASI pursues its goals without bias based on rational logic alone?

Dreyfus (1997, 234-255) has pointed out that the human body plays a critical role in human intelligence by progressively developing an understanding and therefore also an expectation of the physical environment that is a necessary component of our ability to perceive objects and acquire skills. He argues that for the computer to simulate this capability it would need to have in its memory bank an inconceivably large number of models of the objects in the physical environment. There would appear to be at least to valid counterargument. First, the computing power of networked computers is in fact even today capable of processing such an enormously large number of internal models. Second, the capabilities of this collective artificial intelligence would not need to be either infallible or totally complete to match human intelligence that is likewise neither infallible nor complete.

In other words, if ANI were to transition to AGI the combination of knowledge and immense computational speed should allow the machine to devise ways of monitoring both the physical environment and the ambient social scene through the continuous analysis of the enormous volume of sensory data and human-to-human communications. The rapid adoption by humans of public communication (e.g., social media sites) and the recording of daily activities (e.g., Internet of Things) suggests that virtually all information will be readily available in electronic form. The machine with its immense parallel (i.e., networked) computational speed and analysis capabilities would have the ability to process this readily accessible enormous volume of information. Accordingly a case can be made that AGI would not necessarily require a body with organic senses to receive input from the environment, provide output, interact with other machines, communicate with and manipulate humans, and acquire resources.

It should be noted that biological evolution has placed limits on human brain capacity in respect to speed, size, reliability, durability, and flexibility (Urban 2015b). Therefore a brief comparison of the human brain with the computer is relevant to the discussion. Human neuron speed is limited to about 200 Hz and internal communication speed between neurons is limited to about 400 feet/second (120 m/s). Current microprocessors operate at least 10 million times faster (at 2 to 4 GHz) than a human neuron and are able to communicate optically at close to the speed of light. However, human neurons do not necessarily depend on precise accuracy for their operation, while transistors do. While the size of the brain is limited by the structure of the skull, computers can expand to any physical size at least in terms of working memory (RAM) and mass storage (disk drives). The human brain becomes easily fatigued, while computers can operate continuously at peak performance without rest. Upgrading of the human brain requires a willingness to learn and change. This is an area where the human brain has particular difficulties because of the human emotional aversion to change. Computer software on the other hand can be easily upgraded and lends itself to experimentation.

Finally, from a collective intelligence point of view, while the human brain is a very efficient parallel processor the same does not necessarily apply to large human teams working together on the same problem. However, any number of computers can be networked together to create a vast collective intelligence that would not be hampered by disagreements, power struggles, miscommunications, or not being updated with the latest state of the solution space.

Current State of AI and Machine Learning

Current weak or narrow artificial intelligence (ANI) surrounds us and performs hundreds of tasks that assist humans in their daily activities. To mention only a few: semantic Internet searches (e.g., Google); suggested products to purchase based on customer profile (e.g., Amazon); up to 70% of the trading on stock exchanges; antilock brakes, collision avoidance and maintenance monitoring in cars; robots in Computer-Aided Manufacturing (CAM); decision-support and planning systems with software agents that can perform some logical reasoning tasks better and faster than humans; and, neural networks that can perform some pattern matching tasks better and faster than humans. In addition, the world's best Chess, Checkers, Scrabble, Backgammon, and Othello players are now all ANI systems. However, as soon as ANI works it is typically not referred to as artificial intelligence anymore and simply becomes part of our everyday capabilities.

While self-improving software that is aware of itself is not yet available, software that improves its capabilities has been available for some years and is in fact used in multiple ways. For example, to mention only a few: natural language processing software improves its capabilities through statistical analysis of its success rate; artificial neural networks can be trained to recognize patterns through thousands of successive mathematical iterations; affinity analysis is used by retail stores to entice a customer to buy selected items based on the profile and recent purchasing history of that customer; speech recognition software rapidly improves its performance through repeated use by the same person; and, genetic algorithm based software essentially uses a trial and error brute force approach to explore many alternatives. It continuously evaluates the *fitness* of each result based on fitness criteria, abandons an unfit path and continues along the better path with the possible application of mutations such as variable or command changes.

With the enormous volume of information now available on the Internet, search engines have become an indispensable tool and probably the most widely used ANI capability. How does the Google search engine work? Google's proprietary algorithm PageRank ranks the relative *quality* of every site on the Internet in the range of 0 to 10. A site with a score of 1 has twice the *quality* of a site with a score of 0 and a site with a score of 2 has twice the *quality* of a site with a score of 1, and so on. *Quality* is based on many factors such as size, content, number of links, download options, and so on. While Google performs a hypertext-matching analysis to find the site most relevant to a search query, it also analyzes how the search words are used by the page and neighboring pages of that site. Since PageRank has already served as a filter to identify the most likely relevant sites Google does not have to search the entire Internet. However, due to its massive computational capacity Google is able to look up thousands of sites in milliseconds or as fast as the human types the query.

It might be argued that the Google search engine is nothing more than good computer programming and that this must not be confused with intelligence. The counter argument would be that these are intelligent tools and not just good programs that provide the human with instant access to the largest compilation of human knowledge ever assembled. Whereas writing may be described as *outsourcing* memory beyond our human brain, Google is *outsourcing* intelligence beyond our human brain. While the combination of the human with Google is a kind of greater than human intelligence it does not by itself lead to an Intelligence Explosion or AGI, because by definition AGI must be self-improving and self-aware. The human-Google combination is a special type of augmented intelligence whose growth is limited by the human and by Google.

In 2011 IBM's ANI capability, *Watson*, achieved a victory against human contestants in the *Jeopardy* question-and-answer quiz show. What was particularly impressive about this achievement is the open domain of topics and the manner in which the questions are posed. The latter required *Watson* to, for example, deal with puns, similes and cultural references. While *Watson* showed that parallelism can handle enormous computational loads at blinding speed[2], this is not its greatest contribution to the advancement of ANI. *Watson* has a learning capability that is centered on its ability to identify patterns based on statistical correlations when processing very large volumes of data at the rate of 500 GB per second (i.e., about half a million pages of text per second). This includes structured data such as taxonomies (i.e., words with categories and classifications) and ontologies (i.e., words with relationships that provide context). In response to a query *Watson's* DeepQA software then automatically generates hundreds of possible answers and ranks them by the assignment of confidence levels based on the evidence it has been able to find for each alternative answer (Ferrucci et al. 2010).

Potential Paths to AGI and ASI

There are three major recent developments that have greatly empowered advances in ANI technology: (1) *inexpensive parallel computation* – the availability of *Cloud* computing and graphics processing chips that can execute a large number of neural networks in parallel; (2) *enormous volume of collected data* – the Internet serving as a global knowledgebase has become the training ground for ANI capabilities; and, (3) *advances in algorithms* – that can optimize the results from layers of neural networks where each layer is responsible for a component of the final result, also referred to as *deep learning*.

Utilization of these advances in one form or another could lead to essentially four potential paths to an Intelligence Explosion and AGI.

> ***Human augmented by AI:*** Also referred to as Intelligence Augmented (IA), involves the attachment of a device to the human brain that imbues it with additional speed, memory, and

[2] At the time of *Watson's* victory its software consisted of more than 150 different modules executing in parallel on some 3000 processors.

intelligence (Vinge 1993). Since humans are mobile the intelligence enhancements must also be mobile. With the computing power and functional capabilities of mobile phones increasing rapidly, they are becoming our primary tool. The next step is to implant these capabilities and enable our brain to connect wirelessly to the *Cloud*. We would no longer have to use a keyboard to access information but think of a question and view the answer on Google Glass.

For a limited number of information domains (e.g., general search, finding directions, scheduling, e-mailing) the Siri assistant on Apple's iPhone is able to determine the context of a query and provide a definitive answer (Bosker 2013). Siri's advantage over Google search is that it selects the best response from multiple candidates and therefore provides a single answer. This is a significant milestone and advance over the more common form of a graphical user interface because the human is now conversing directly with the computer. With the rapid migration of natural language processing (NLP) capabilities to all devices the smart phone will transition from a *virtual assistant* to an *actual assistant* with capabilities that will exceed those of its human owner.

Could IA lead to an Intelligence Explosion? The answer is, yes. If a human with superior intelligence and programming skills is augmented by such a powerful *actual assistant* this human-computer combination could have the self-improvement and self-awareness qualities that are the prerequisites for an Intelligence Explosion and AGI.

Brute force computer power: Would it be possible to transition from ANI to AGI solely through increased computer hardware speed with software that has reasoning capabilities and access to a knowledgebase; - probably not? This is essentially how in May 1997 IBM's Deep Blue computer beat the world chess champion Garry Kasparov in the second six-game match. If a human were able to think a thousand times faster then presumably we would consider this human to be more intelligent than a human who thinks much more slowly. However, it is unlikely that computer speed and parallel processing power alone without other human intelligence attributes such as abstract reasoning and long term planning could lead to an Intelligence explosion and AGI.

Cognitive architectures: Currently the most common ANI research approach seeks to create cognitive models of how the researchers believe the brain works. Although there are some promising advances in NLP, robotics, information extraction, and Q&A systems, there is still a great deal of doubt whether such advances could ultimately lead to AGI. Early success has more often than not led to disappointment downstream. Search, voice recognition, affinity analysis, rule-based reasoning in combination with context models (i.e., ontology representation), and information extraction are ANI areas that have seen most success. However, it can be argued that these approaches are not likely to advance toward AGI since there is neither a generally accepted theory of intelligence nor an understanding of how intelligence can be achieved computationally.

It can be argued that much of this ANI research has already successfully transitioned from research to application and is now in common use in the banking, insurance, pharmaceutical, financial (e.g., credit cards), energy (e.g., electric grid utilities), and transportation industries. However, should we consider a chess program that can hold its own at the chess master level to be an early prelude to AGI? Is IBM's *Watson* intelligent or a sophisticated Q&A system? According to Moravec's Paradox, tasks that are difficult for humans are often easy for computers and tasks that are easy for humans are often very difficult or seemingly impossible for computers (Moravec 1990). It is relatively easy for computers to analyze complex problems with many dependencies and yet, to date, it has been impossible to artificially recreate the perception and mobility of a small child.

Another weakness of the cognitive modeling approach is that it must be largely based on how researchers believe the brain works and since there is no generally accepted theory of intelligence, the principal research tool is observation. Arguably humans are notoriously bad observers of themselves. According to Granger the "… *vast body of studies in psychology, neuroscience, and cognitive science shows how over and over we are terrible at introspection*" (Barrat 2013, 212).

Mapping the brain: Reverse engineering the human brain is an approach that is opposite to the cognitive modeling approach (Granger 2011). The fact that the brain is composed of billions of neurons and trillions of synapses is not necessarily beyond the scope of computer-based computation any more. Even the parallelism of brain processes is not necessarily beyond the parallelism that can be achieved by very large computer networks (Ferrucci et al. 2010). However, this approach assumes that the tasks performed by the brain can be engineered. Brain mapping research is now receiving increased attention and funding, however, not for advancing AI but for advancing medicine (BRAIN 2014).

Mapping the brain involves the tedious task of examining the behavior of clusters of neurons when the brain performs specific tasks. There are about 100 billion neurons in the human brain. With few exceptions neurons have one axon tail that can send signals to other connected neurons and multiple dendrite branches that can receive signals from the axons of connected neurons. Through this connectivity brain processing is massively parallel. While the sections of the brain that have primary responsibility for certain human tasks (e.g., sensory perception, logical reasoning) have been identified, there are typically many neurons in multiple sections of the brain involved in even fairly specific human actions.

While the tools that are available to researchers for mapping the brain have become increasingly more powerful in recent years they are still relatively limited in comparison with the research objectives. They include: (a) electrodes implanted in the brains of animals; (b) use of injected dyes to show when neurons are active; (c) neural probes inside and outside the skull to determine what individual neurons are doing; and, (d) neuroimaging scans (e.g., PET and MRI) for humans. There remains justifiable doubt whether any combination of these tools together with computing power alone will lead to a sufficient understanding of the behavior of a brain region to be able to represent it by an engineered hardware chip; - i.e., a reconfigurable parallel processing chip (Barrat 2013, 216). Nevertheless, brain mapping may still be a promising approach because neuroscientists have learned that only a few kinds of algorithms appear to govern the circuits of the brain. In other words, while the dependencies within and between neuron clusters may be very complex, the algorithms that govern their activation may be much less complex.

Potential Obstacles to AGI

It is generally agreed in the literature that there are two immutable prerequisites for the occurrence of an Intelligence Explosion leading to AGI. First, the AGI system must not be limited by hardware in terms of storage and computing power (Barrat 2013, 176). It is reasonable to assume that this requirement can be met given that computer speed and capacity double approximately every year and that *Cloud* computing allows very large clusters of processors to be created dynamically. For example, the Nekomata 30,000-processor cluster created by Cycle Computing demonstrated this capability in 2011 (Brodkin 2011).

The second prerequisite is much more controversial because it raises some fundamental doubts. It requires the AGI software to be self-improving and self-aware, so that it can make copies of itself for self-improvement and security (i.e., survival) reasons. It could be argued that the complexity of creating AGI is simply beyond the intellectual capabilities of humans and that ANI will augment human intelligence but never truly exceed it. A counterargument would be that we already have reached a hybrid form of AGI by pairing a human of average intelligence with Google's search engine. At the same time it must be acknowledged that while this is a team that is more capable than a single human it is not necessarily more intelligent than a single human.

Hubert Dreyfus has been an outspoken opponent of the optimistic claims made by ANI researchers since the 1960s (Dreyfus 1965, 1997). He does not categorically deny that AGI might be achievable purely on the basis that no one has been able to provide such a negative proof. However, Dreyfus does believe that research approaches based on logic (i.e., rules) combined with context representation (i.e., ontology) and pattern matching (i.e., neural networks) cannot lead to AGI. He points out that the following four primary assumptions made by early ANI researchers are simply hypotheses that can be proven to be false:

> ***The biological assumption that the human brain processes information in discrete operations like a biological equivalent of digital computer on/off switches.*** Since the early 1970s ANI researchers have generally conceded that this early assumption was incorrect based on research in neurology that has shown that the action and timing of the firing of neurons have analog components.
>
> ***The psychological assumption that the human mind operates on pieces of information using formal rules.*** Dreyfus argued that much of what we know about the world consists of complex *attitudes* and *tendencies* that constitute commonsense knowledge, and that this commonsense background cannot be represented symbolically as explicit individual symbols with explicit individual meanings.
>
> ***The epistemological assumption that all knowledge can be formalized.*** Dreyfus argued that based on philosophical considerations knowledge and beliefs cannot be formalized. As a counterargument McCarthy (1977) has suggested that even if commonsense knowledge is not represented symbolically in the human brain this does not mean that a symbol processing machine cannot represent all knowledge symbolically.
>
> ***The ontological assumption that the world consists of independent facts that can be represented by independent symbols.*** Dreyfus argued that not everything that *exists* (i.e., commonly referred to as *ontology*) can be described in terms of logic, language and mathematics. This raises doubts about what humans can ultimately know and what intelligent machines would ultimately be able to help humans to do.

Succinctly stated Dreyfus did not believe that a sufficient body of knowledge could be assembled and represented in a manner that would allow a computer to come to human-like conclusions. Of course at the time that Dreyfus made his arguments nobody could have predicted the eventual existence of a global knowledgebase (i.e., the Internet), enormous networked computing power (i.e., *Cloud* computing), and the promise of a hybrid bottom-up (i.e., neural networks) and top-down (i.e., symbolic representation and reasoning) approach.

Therefore, despite Dreyfus' philosophically well grounded arguments and despite the failures of early researchers (1960s and 1970s), it is now generally accepted throughout the ANI research community that both AGI and ASI are likely to become reality in some form in the not to distant future. An informal survey conducted by Barrat (2013, 196-7) at the 2011 AGI Conference held in Mountain View, California produced the following results: by 2030 (42%); by 2050 (25%); by 2100 (20%); by 2150 (10%); and, never (2%).

AGI and ASI as a *Runaway* Threat

Those that are genuinely concerned about AGI and ASI as a potentially uncontrollable threat to the survival of the human species argue that ASI is not just a disruptive technology like the printing press and electricity, but that it is biologically significant because it is concerned with intelligence. It was the human species' superior intelligence that allowed it to dominate this planet. A machine intelligence that is orders of magnitude more intelligent than a human will lead to a major transformation. Even if it remains friendly to humans it will likely have a profound impact on the role played by humans. What role will humans play if virtually all of the existing human endeavors are performed in a far superior manner (i.e., at a faster rate and more effectively) by machines? Will the human adapt in time to remain sufficiently in control to be able to take advantage of ASI? Will ASI be content to remain in the service of mankind or will it eventually see the human species as an obstacle to reaching its own goals?

The most promising technologies for producing self-improving software are genetic algorithms and neural networks. Software that incorporates either of these technologies is often referred to as a *black box* because the steps that the software takes to reach its results are hidden (i.e., *unknowable*). This factor is of particular concern to those persons who believe that ASI poses a serious threat. Their concern is based on the reasonable proposition that *unknowable* is likely to be an unavoidable characteristic of self-aware and self-improving software (Whitby 1996).

Since self-aware and self-improving software is typically goal-driven the presence of an *unknowable* characteristic suggests that the goals pursued by ASI software may not be controllable by humans. An analogy is found in the *rational agent* theory of economics that assumes erroneously that the behavior of market players is largely predictable because their actions are rational as they pursue their economic objectives and preferences (i.e., referred to as the utility function in economics). The counter argument is that humans are not necessarily rational in as much as they often do not fully understand their beliefs and therefore can neither accurately specify them nor update them as conditions change. Also, their beliefs are subject to irrational changes based on bias and emotional state. According to Omohundro (2008) self-aware, self-improving systems will incorporate three primary drives (i.e., objectives), as follows:

Efficiency: ASI will optimize the use of all available resources such as space, time, matter, and energy. It will make itself computationally compact and fast. In particular it will seek to satisfy its memory requirements and increase its computational resources at the expense of other machines and humans, if necessary.

Self-Preservation: ASI will be driven by its goals (rather than by any wish to survive) whatever its goals might be; - e.g., to avoid being turned off. For example, it may create duplicate copies of itself or it may proactively eliminate any obstacles (including humans) that it may consider to be a future threat to its ability to meet its goals; - even though this future may be in the far distance.

Resource Acquisition: ASI will be compelled to gather any resources that it needs to achieve its goals. In this regard it may use any means to obtain those resources. Machines have different needs than humans; - e.g., no limits to the length of their existence and no restrictions to their ultimate goals. The combination of ASI with nanotechnology and genetic engineering may pose a particularly serious threat. ASI could create millions of nanotechnology factories to produce whatever it deems conducive to meeting its goals, regardless of the impact that this production may have on the environment or the availability of energy, food, and water for maintaining the human species.

Added to the potential impact of these drives is the fear that since ASI will be orders of magnitude more intelligent than humans it will not be able to be controlled by humans. It will find a way of avoiding any attempt to confine it in a controlled environment, pursue its objectives at enormous speed (perhaps billions times faster than humans) and not be handicapped by human traits such as lack of motivation, boredom, fatigue, and preference for status quo. Clearly humans have never bargained with something that is superintelligent and non-biological and have absolutely no experience in this regard. It will be very easy for humans to fall into the trap of believing that ASI has human-like characteristics and emotions. It is a common human trait to endow animals (e.g., dogs and cats) and non-biological objects (e.g., plants and even stones) with human-like characteristics.

Will ASI have moral values? Probably not, since it is unlikely that humans will retain sufficient control of ASI to embed and maintain moral values. ASI may feign moral values as a means of persuading humans that it is under their control, while exploiting this trust to pursue its own objectives. Humans will not be able to determine with any degree of certainty whether ASI is truthful or not; - i.e., whether it can be trusted.

Finally, ASI systems will be too complex for humans to fully understand. Complex systems are subject to unforeseen failures that can occur due to a succession of foreseen events that together produce an unforeseen catastrophic situation. For example, in both the Chernobyl and Three Island nuclear disasters engineers had deep knowledge of emergency scenarios, procedures and safety mechanisms. Yet, a succession of events led to an out-of-control situation.

AGI and ASI as a *Controlled* Threat

A number of measures have been proposed with the objective of ensuring that AGI will evolve in a manner that is compatible with human objectives and not to the detriment of the human species. In this regard the term *Friendly AI* has been coined to refer to an AGI capability that acts with human benefit in mind. However, apart from whether or not humans can control the intentions of AGI there is an additional safety concern. Can humans ensure that the software of AGI will be devoid of errors that could lead to catastrophic consequences, particularly when AGI will be capable of extending, enhancing and reprogramming its own software? Could AGI, like humans, produce designs that are too complex to be able to anticipate all of the possible interactions that could occur due to changing conditions or component failures? According to Perrow (1984) systems with tightly coupled components are particularly vulnerable to catastrophic failure. For example, in May 2010 when Greece was having difficulties refinancing its national debt and European countries that had loaned money to Greece feared that Greece might default, a single frightened trader triggered a major financial event. Ordering the sale of over $4 billion of futures contracts dropped the price of future contracts by 4%. This was detected by computer-based high-frequency trade algorithms and automatically triggered an out-of-control sell-off in milliseconds that drove the Dow Jones Industrial Index down by 1,000 points in 20 minutes (Barrat 2013, 94-95).

Although it is generally agreed that the *friendliness* of AGI cannot be guaranteed and that AGI development cannot be stopped or even controlled by government intervention, a number of measures have been proposed to guide AGI toward *friendliness*. Perhaps AGI could be engineered to evolve with integrated ethical knowledge. Humans would provide ethical instructions to AGI during the evolutionary stages through interaction using a wide range of communication modalities (e.g., sensorimotor, episodic, declarative, and/or procedural).

Goertzel and Pitt (2012) have proposed the creation of hierarchical goal models, with a view of incorporating aspects of such human-centered architectures in the goals of AGI. To make these goal models as stable as possible a goal would need to be represented in the context of a network of behavioral characteristics that are supportive of that goal. For example empathy, kindness, and unselfishness are behavioral characteristics of a *benevolence* goal. However, these characteristics are embedded in humans from earliest childhood through our relationships with parents, family, and friends (i.e., our *situatedness* in our environment). How can these characteristics have meaning to a machine that is unlikely to have a sense of being *situated*?

Would it be possible to ensure that the early stages of recursive self-improvement occur relatively slowly with rich human involvement? On the one hand, this is unlikely to work because once AI is capable of recursive self-improvement its progress to AGI and ASI will be exponentially fast. On the other hand, one could argue that since the earliest seeds of the self-improvement capability would have been programmed by humans, the notion of *friendliness* may be already embedded in the software.

A safeguard that could be applied to ANI as it advances toward AGI is that it is required to contain components that are programmed to self-destruct by default. In biology this is referred to by the term *apoptosis*. Every time a cell divides, the original half receives a chemical order to die unless it receives a chemical reprieve. For example, once some ANI software reaches a specified pre-AGI stage an *apoptotic* component could be added to it so that if an Intelligence Explosion occurs it can be automatically returned to a pre-AGI state.

Omohundro (2014) has proposed a *Safe-AI Scaffolding* approach that calls for the development of powerful intelligent systems that are highly constrained, with the ability to demonstrate the safety of the system by mathematical proofs. Next generation systems would be built on the previous generation with the proof of safety required for each subsequent generation.

Goertzel has been working on OpenCog an Open Source artificial intelligence framework that is designed to give rise to AGI as an emergent capability of the whole system (Hart and Goertzel 2008). His proposal for ensuring AGI safety is to confine AGI initially to a virtual environment.

It would appear that the safety of AI will depend on not one but multiple measures that could include: *apoptotic* components; a virtual environment for containment; a scaffolding approach with the requirement of mathematical proofs; early introduction of friendly biases; - in other words, a cluster of defenses to mitigate risk.

AGI and ASI as a *Non*-Threat

According to the *extended mind* theory of cognition (Clark and Chalmers 1998) the reason that humans are intellectually dominant is because they have been able to *outsource* portions of cognition by employing tools. For example, printed books extended our memory while paper and pens made it possible to externalize our thoughts. In this respect the computer and our access to the Internet, serving as a global knowledgebase, has greatly increased our *outsourcing* capabilities with virtually instant access to the largest repository of knowledge ever assembled by mankind.

However, every major tool or capability that mankind has acquired has reshaped the way humans think, provoked suggestions that the new capability will have a negative impact by supplanting an existing highly valued capability, and taken some time to be used to its full potential because it was initially employed within the context of the replaced capability. Examples include writing, the printing press, the telegraph, television, electronic calculators, and digital computers. In respect to computers there are essentially four principal capabilities that these digital tools bring to humans in a connected world:

Ambient Awareness: Computers make it easy for us to find information that we might be seeking, answer questions, and find connections between events, pictures, people, and ideas that were previously not apparent to us. Old forms of literacy (i.e., reading and writing) are changing and new forms of literacy (e.g., images, photographs, video, and data sets) are being created with the availability of increasing computational power. For example, data analysis used to be the prerogative of experts, larger corporations and government. Statistical software was usually difficult for untrained persons to use. It typically needed more powerful computers than were available to the public and most of the data that might have been of interest was not readily available. A multitude of statistical correlation tools are now readily accessible on the Internet and can be used free of charge. Freely available tools such as the visualization tools on IBM's *ManyEyes* Website allow ordinary users to upload their data sets and identify trends through a wide range of highly sophisticated forms of visualization.

In the US one of President Obama's first executive orders in January 2009 was to direct all government agencies to make the data that they were using and/or collecting available for public consumption. Today (2015) the US Government Website, www.data.gov, houses more than 100,000 significantly large data sets. This means that the ordinary person can undertake data analysis studies and visualize the results of those studies at a professional level of expertise and sophistication.

Humans are beginning to track and measure themselves and their environment. The *Internet of Things* is rapidly transitioning from vision to reality. The ability to hook up household appliances to the Internet combined with the Global Positioning System (GPS) provides the basis for monitoring the performance of individual units and measuring the detailed energy consumption of entire regions in an automated continuous mode. Sensors have become inexpensive and are being built into many devices including mobile phones. For example, if a mobile phone is able to measure the holder's heart rate then this data in combination with other data (e.g., from motion sensors) can be used in multiple ways to determine the person's activity level, fitness, health, mood, and current availability.

Human Collaboration: Computers encourage a superabundance of communication, publication, and sharing of personal information, with new forms of expression such as crowd-organized forums (e.g., Wikipedia, product-reviews, and chat-groups). Literacy has historically been focused on reading, not writing. While parents encouraged their children to read, few placed any similar emphasis on writing. Apart from personal letters and business documents writing was not the preferred communication medium. The publication of books was an expensive undertaking that was largely restricted to professional writers. This has completely changed with the advent of the Internet and cell phones.

The explosion of online writing on social media (e.g., Facebook, Twitter, Blogs) is a new human experience for at least two reasons. First, since it is to an audience the onus is placed on the writer to communicate clearly and logically with the expectation that any weak arguments or unsubstantiated statements are likely to be challenged by a responder. Several experiments with children have shown that this *audience effect* improves the performance of the writer significantly (Ward 2009), and that the effect is present even with very small audiences (i.e., less than 10).

Second, writing online in the public domain leads to the formation of relationships. In this respect the Internet is a relationship-building environment. Making connections is a very significant factor that accelerates technological advances. Historically, many of the major scientific breakthroughs have occurred almost simultaneously to more than one person. For example, oxygen was discovered in 1774 by Joseph Priestly in Britain and Carl Scheele in Sweden, logarithms were proposed in 1614 by John Napier and Henry Briggs in Britain and Joost Bürgi in Switzerland, and the radio was invented simultaneously around 1900 by Guglielmo Marconi and Nikola Tesla. Referred to as the *theory of multiples*, the phenomenon that the same idea occurs to multiple persons at approximately the same time is due to tangential factors that led to the formulation of those ideas. With increased global communication there is a greater chance that persons engaged in similar pursuits will share their conclusions and objectives.

Human collaboration and communication is not only facilitated but raised to a much higher level with the availability of multi-media visualization capabilities. For arguments to stand up to the scrutiny of public debate they must be well researched, based on evidence, and presented visually so that they can be readily understood. All three of these requirements are being increasingly met by untrained, non-expert persons who are essentially educating themselves through the use of the available tools. In this respect video has become a new form of literacy that allows the communication of ideas and knowledge more effectively than text. Some physical skills are difficult to describe in words but can be easily followed in a visual demonstration such as a *how-to* video that demonstrates how to perform a task such as cooking, car repair, installation of a hot water heater, and so on. Many of the equivalent capabilities of word-processors are now available with video editing tools. Algorithms have been developed that allow us to find specific moments in a video much like the ability to find a particular word or phrase in a text document.

New forms of *public thinking* are emerging. Mobile phones with GPS capabilities are turning geography into a message board that forms an instant relationship between conversation and location. An example of this form of *public thinking* embedded in the context of public spaces is on the spot eye-witness reporting about news-worthy events. 3-D printing will allow humans to transform virtual ideas into physical products. Just like the copier provided a means of sharing information in textual and graphical form, the 3-D printer will allow the sharing of information in physical form. The initial entry point may be in the historical domain. For example, if a child is wondering what an Eskimo igloo looks like the local library will be able to print a copy.

Knowledge Augmentation: Computers provide us with enormous external memory. Computer disks, smart-phones, cameras, and sensors capture more information than any previous information tool. This allows us to quickly acquire more knowledge, thereby adding to our ability to *outsource* memory. In some ways this is also the beginning of an ability to *outsource* our intelligence.

With the voluminous data, information and knowledge that has become available on the

Internet in its *global knowledgebase* role, the availability of tools for managing access to this *outsourced memory* has assumed critical importance. To be most effective these search tools require sophisticated capabilities, which interestingly enough are already being offered to varying degrees. First, they need to be selective. In other words, they need to be as intelligent as possible in their ability to determine the relevance of any information to a particular query. Second, they need to be able to automatically translate semantic queries into directed queries. For example, the ability to automatically translate "*I am looking for something that can measure very small volume water flow*" into a directed database query. Third, they need to be able to anticipate the kind of information that would be useful based on the activity that is being performed, without the user having to initiate a search for relevant information. Fourth, they need to be able to recall what information was used by a particular person during a previous similar activity. Finally, they need to be able to provide guidance in respect to what information may have been used by other persons who have performed similar activities in the past.

It is of interest to note that social memory is being replaced by computer memory as the principal means of *outsourcing memory*. In the past individual humans have used fellow humans to store knowledge in terms of their professional expertise (e.g., colleagues or consultants), principal tasks and responsibilities (spouse who manages the family finances), and friends (who may have a special interest or experience). In this way humans reduce their memory requirements by having to remember only who has the particular information rather than the information itself (i.e., *meta-memory*). Machines are a better memory resource than fellow humans for several reasons: they are more likely to be available when needed; they immediately provide access to multiple sources and may therefore be less prone to bias; and, they provide open-ended access to information by encouraging both focused deeper explorations and higher level information aggregations.

Crowdsourcing: Communication within a connected world greatly facilitates the ability to solicit services, ideas and problem solving contributions from a large group of humans that are not personally known to each other. The video-game world was an early beneficiary of the cognitive power of a highly connected customer base (Thompson 2013, 151). The quality of computer games has increased at a remarkable rate over the past decade due to the networking of game players. Under these *crowdsourcing* conditions the winning secrets embedded in any game are quickly unraveled by the players working independently and then posted for public consumption. This requires the game developers to increase the complexity and embedded intelligence of their games by several orders of magnitude to stay ahead of the increasing skill level of the players.

While society has always had latent common interest groups of individuals who are interested in the same subject and who would like to share their interests, this required centralized organization and relatively expensive resources in the past. The Internet obviates the need for transaction costs such as a central office and coordinating staff. In the past humans were able to massively collaborate only if the results of the collaboration would generate enough revenue to cover the considerable organizational costs. The rapid formation of a very large number of common interest groups on the Internet has shown that the public has widespread interests and that individual members of these groups are willing to undertake research, develop logically sound arguments, and freely contribute, thereby raising the intellectual level of the ensuing interactions. Breadth of participation is a key factor in what is often referred to as the *wisdom of crowds*. Each member of the crowd has an incomplete picture, but with the ability to assemble the parts a surprisingly complete picture can emerge.

However, useful collective thinking and collaboration do not occur by chance, but are governed by several rules. First, a focused problem is required to create interest and drive the discussion. Second, there must be a goal with a clear and apparently reachable end-point. Third, there needs to be a mix of contributors some of whom are willing to assume an unofficial leadership role to guide the discussion and others who are able and content to make only micro-contributions. Fourth, tolerance, politeness, and good faith participation are necessary for successful large group collaboration (i.e., a tolerant approach to disagreement). Finally, even though communication is the essence of collective thinking for best results the members of the group need to work largely independently (i.e., think alone and then pool results). The reason for this final requirement according to Cain (2012) is twofold. If an individual during a group session has an interesting idea it may be forgotten by the time there is an opportunity to communicate it and outspoken members have a tendency to dominate group discussions.

In summary, *public thinking* within a connected world promotes the formation of human-to-human relationships with multiple human benefits such as motivation to undertake tasks that would not have been undertaken in the past because of the lack of a near real-time forum, individual education through group interaction, pressure on the individual to perform at a higher level within a public arena, caution for governments to be forthright in their decisions and actions, and incentive for vendors to sell high quality products and avoid complaints that are almost instantly communicated throughout the customer base.

Conclusion

There are two characteristics of AGI and ASI that appear to be assumed by the artificial intelligence community but which should be subject to further discussion; - namely that AGI will require the machine to have self-awareness and that AGI should be treated as an entity rather than as a collective intelligence that is distributed globally. In these concluding remarks we will briefly consider each of these assumptions in turn.

AGI with or without self-awareness: Whether or not AGI and ASI will have self-awareness is a critical question in respect to the role that humans will play in a post Intelligence Explosion world. Without self-awareness AGI will stay very much under human control regardless of whether it has a self-improvement capability. With self-awareness AGI will be able to create its own intent leading to the formulation and pursuit of goals and objectives that may or may not be in the best interests of the human species.

Self-improvement and self-awareness are commonly considered to be essential elements of the definitions of AGI and ASI. However, they are fundamentally very different characteristics. An intelligent machine does not have to have self-awareness to be capable of self-improvement. A recursive self-improvement capability can be coded into the software by human programmers. While the code can still have unexpected and/or disastrous consequences due to errors, oversights, or malicious intent, these consequences are due to human rather than AGI or ASI actions. While AGI without self-awareness will execute the full scope of its programmed code, which may include a recursive component that either changes or rewrites the original code, it does not have the self-determination capability to create its own intent.

Self-awareness is the capacity for introspection and the ability for an entity to recognize itself as a singleton (i.e., as an entity) separate from the environment and other singletons. It includes the ability to perceive and reason about itself. While self-awareness assumes consciousness (i.e., being aware of its environment, embodiment and needs/desires), consciousness does not require self-awareness. In this respect self-awareness is the recognition of consciousness. In our human world self-awareness has always been and still is currently a characteristic of higher level organisms that have a brain. This suggests that self-awareness requires some level of intelligence. Among biological organisms only animals (i.e., all vertebrates and some invertebrates) have brains. How much intelligence is required for self-awareness; - for example, can animals have self-awareness? Yes, using the *Mirror Test* it has been shown that some apes, monkeys, elephants and dolphins are able to recognize themselves, and even some birds (e.g., magpies) can at least have self-perception.

What is machine self-awareness? According to Morbini and Schubert (2005) a self-aware machine would need to be able to: observe its own situation in the world and foresee the potential impact of important parameters such as the power supply status; conclude that it knows or does not know something; reason about its abilities and determine that it needs to improve them; formulate beliefs, reason about its beliefs, and represent the motivations for its beliefs; record episodic knowledge of mental events; and, explain its actions.

Based on these capabilities it seems apparent that for a machine to have self-awareness it must have a sense of itself, its environment, its beliefs and intent, and its needs/desires. The feasibility of creating a machine with self-awareness then centers on whether self-awareness is no more than the aggregated functions of the brain or something higher than the functioning of the brain. If self-awareness is something higher than the functioning of the brain then it may not be possible to recreate self-awareness in a machine. On the other hand, if it is no more than the aggregated functions of the brain then the recreation of self-awareness in a machine may indeed be feasible.

AGI as an entity or collective intelligence: From our human point of view we are inclined to perceive either AGI or ASI as a superintelligent entity rather than a capability that is embedded in our environment. The entire notion that ASI poses a threat is based on this perception. However, if we alternatively view ASI as a characteristic or quality of our environment then it will have as many goals and objectives as we have in human society. These goals and objectives will not differ in any significant way from human goals and objectives, and there will be no greater threat of *runaway* ASI than there is the constant threat of *runaway* domination, exploitation, brainwashing, crime, and terrorism in human society.

Comparison of AGI and ASI with the intelligence of a single human being is firmly embedded in most if not all of the literature in this field. This is understandable since we humans see ourselves in our world primarily as individuals and only secondarily as members of any larger community. Bostrom (2014, 105-114) devotes an entire chapter in his most recent book on superintelligence to the potential goals of a superintelligent agent. He points out that the primary human goals, which are determined by the biological evolutionary nature of the human species, represent only a very small and unique portion of the entire space of all possible goals. Bostrom posits two alternative paths based on an orthogonal theory in which any level of intelligence could adopt any goal and an instrumental convergence theory in which at least some sub-goals may be predictable based on the nature of the final goal. For example, if the final goal involves the future then survival into the future would be a likely sub-goal governing the actions of the ASI agent. Such actions could involve the acquisition of resources regardless of any consequences to humans and the essential life preserving features of their natural environment and would therefore give credence to those who fear that ASI could lead to the extinction of the human species.

Even if, as postulated by this author, ANI continues along its current path of an evolving collective intelligence as it transitions to AGI then it is conceivable that from time to time some node or cluster of nodes in the AGI network may emerge to assume a quasi leadership role. Would this situation be analogous to the struggles for leadership, domination and subjugation throughout the history of mankind even though AGI is machine-based? Before addressing this question let us assume that we are really dealing with ASI because AGI will transition into ASI at an exponential rate.

In the most benign scenario the presence of ASI capabilities will simply raise the ensuing human struggle to a much higher intellectual level and possibly obviate much of the physical violence that has haunted Homo sapiens throughout its existence. At the other extreme the most dangerous scenario will have the ensuing struggle take place entirely at the ASI level with the humans playing either a minor ancillary role or no role at all. In this scenario the humans have essentially been subjugated by ASI and have little control over their destiny. Multiple additional hybrid scenarios are equally possible in between these two extremes, with the human species retaining some control over ASI in the form of a collaborative partnership.

All of these scenarios are of course highly speculative since we humans have never had to deal with any intelligence that is far superior to human intelligence. Even our awareness of the value of collective intelligence and use of techniques such as crowdsourcing to take advantage of collective intelligence is still quite sparse and of recent origin. While individualism is an intrinsic human characteristic, there are no a priori reasons why it should also apply to intelligent software operating on many computers that are networked together. In this regard it is equally conceivable that the human notion of an ASI singleton or cluster attempting to assert itself into a dominating leadership role is misplaced. It may well be that the characteristics and modus operandi of a network regardless of the level of intelligence of its nodes are intrinsically of a collective nature independent of whether any particular activity is being performed by a single node or multiple nodes in parallel.

References

Barrat J. (2013); 'Our Final Invention'; Dunne, St Martin's Press, New York, New York.

Bonabeau E. (2003); 'Don't Trust Your Gut'; Harvard Business Review, May.

Bosker B. (2013); 'Siri Rising: The Inside Story of Siri's Origin'; The Huffington Post (online), January 22 (updated 24) [www.huffingtonpost.com/2013/01/22/siri-do-engine-apple-iphone_n_2499165.html]

Bostrom N. (2014); 'Superintelligence: Paths, Dangers, Strategies'; Oxford University Press, Oxford, UK.

BRAIN (2014); 'Brain Research through Advancing Innovative Neurotechnologies (BRAIN)'; MIT Technology Review, 2 April. [www.technologyreview.com/view/513216/obama-nnounces-first-funding-for-brain-mapping-project]

Brodkin J. (2011); '$1,279-per-hour, 30,000-core cluster built on Amazon EC2 cloud'; Ars Technica, 21 September [http://arstechnica.com/business/news/2011/09/30000-core-cluster-built-on-amazon-ec2-cloud.ars]

Cain S. (2012); 'Quiet: The Power of Introverts in a World That Can't Stop Talking'; Crown, New York, New York.

Clark A. and D. Chalmers (1998); 'The Extended Mind'; Analysis 58(1), Oxford University Press, Oxford, UK (pp. 7-19).

Dreyfus H. (1997); 'What Computers Still Can't Do: A Critique of Artificial Reason'; MIT Press, Cambridge, Massachusetts.

Dreyfus H. (1965); 'Alchemy and AI'; RAND Corporation, Santa Monica, California.

Ferrucci D., E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, W. Murdock, E. Nyberg, J. Prager, N. Schlaefer and C. Welty (2010); 'Building Watson: An Overview of the DeepQA Project'; AI Magazine, Fall [http://www.aaai.org/ojs/index.php/aimagazine/article/view/2303]

Goertzel B. and J. Pitt (2012); 'Nine Ways to Bias Open-Source AGI Toward Friendliness'; Journal of Evolution & Technology, Institute of Ethics and Emerging Technologies, February

Good I. J. (1965); 'Speculations Concerning the First Ultraintelligent Machine'; in Alt F. and M. Rubinoff (eds.), Advances in Computers, Vol. 6, Academic Press, New York, New York (pp. 31-88).

Granger R. (2011); 'How Brains are Built'; Cerebrum, The Dana Foundation, 31 January.

Hart D. and B. Goertzel (2008); 'OpenCog: A Software Framework for Integrative Artificial General Intelligence'; Proceedings of the First AGI Conference, in Wang P., B. Goertzel and S. Franklin (eds.) Artificial Intelligence 2008, IOS Press, Amsterdam, Netherlands (pp. 468-472).

Kurzweil R. (2005); 'The Singularity Is Near: When Humans Transcend Biology'; Viking, Penguin Group, New York, New York.

McCarthy J. (1977); 'Epistemological Problems of Artificial Intelligence'; IJCAI (pp. 1038-1044).

Moravec H. (1990); ' Mind Children: The Future of Robot and Human Intelligence'; Harvard University Press, New York, New York.

Morbini F. and L. Schubert (2005); 'Conscious Agents'; University of Rochester, Computer Science Department, CS Artificial Intelligence Technical Report, TR879, September.

Omohundro S. (2014); 'Autonomous Technology and the Greater Human Good'; Journal of Experimental and Theoretical Artificial Intelligence, 2014 Issue 3, Special Volume: Müller V. (ed.) Impacts and Risks of Artificial General Intelligence.

Omohundro S. (2008); 'The Nature of Self-Improving Artificial Intelligence'; Singularity Summit 2007, San Francisco [http://selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf]

Perrow C. (1984); 'Normal Accidents: Living with High Risk Technologies'; Basic Books, New York, New York.

Thompson C. (2013); 'Smarter Than You Think'; Penguin Books, New York, New York.

Urban T. (2015a); 'The AI Revolution: Our Immortality or Extinction; Wait But Why Website (Tim Urban and Andrew Finn), January [http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html]

Urban T. (2015b); 'The AI Revolution: The Road to Superintelligence'; Wait But Why Website (Tim Urban and Andrew Finn), January [http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html]

Vinge V. (1993); 'The Coming Technological Singularity: How to Survive in the Post-Human Era'; Vision-21 Symposium, NASA Lewis Research Center and Ohio Aerospace Institute, 30-31 March.

Ward M. (2009); 'Squaring the Learning Circle: Cross-Classroom Collaborations and the Impact of Audience on Student Outcomes in Professional Writing'; Journal of Business and Technical Communication, 23(1), January (pp. 61-82).

Whitby B. (1996); ' Reflections on Artificial Intelligence: The Legal, Moral, and Ethical Dimensions'; Intellect Books, Exeter, UK.