Artificial Intelligence (AI) — The Need for New Safety Standards and Methodologies

by Malcolm Jones
Reading, U.K.

There have been a series of challenges in developing appropriate safety standards and methodologies as technology evolves to ensure their safe implementation. These challenges, which first arose at the dawn of the industrial revolution, will inevitably continue. New technologies will always forge ahead in a competitive marketplace; failure to do so will inevitably lead to organizational demise. However, these developments must be matched by a complement of research activity seeking to ensure that appropriate new safety standards and methodologies are put in place to maintain acceptable levels of risk. A new challenge now confronts us in the form of artificial intelligence (AI), where we stand at the frontiers of decision making in relation to what roles machines and humans should play in optimal decision making and how this will impact safety.

Introduction

We are all now aware of the now well established implementation of automation in industrial processes with its benefit in relation to removing and/or reducing the load on humans, together with greater control of quality and efficiency in product manufacture, especially in the context of mass production and safety. By and large, these processes are set up and controlled by humans via "hardwired" or software-controlled methodologies. These implementations required the parallel development of safety methodologies and relevant safety standards together with associated regulation to ensure appropriate levels of safety compliance. Of course, the safety standards and their required level of compliance will be related to the level of consequence if an inadvertent event (mishap) occurs. If this lies in the "catastrophic" range, which very much applies to the subject of nuclear warhead design, assembly and ownership, then the requirements will be exceptionally onerous. It is for this reason — and not surprising — that such processes have not yet been automated in the assembly context but have in the context of individual components manufacture. Such components may not be unsafe in themselves, but are nevertheless subject to intense human scrutiny prior to "incorporation" in the assembly process where their safety attributes can be paramount.

However, the general field of automation is now developing rapidly and into areas where the control methodologies are becoming very complicated and sometimes difficult for humans to fully visualize or understand. By extension, this is moving to the level where decisions, and hence control actions, are being taken out of human hands. That is handed over to machines via Artificial Intelligence (AI). AI can have advantages when it exceeds human capability and where the information available may be too complex for humans to handle or understand or respond to in sufficient time.

Currently, there is no clear vision of the rate of development or where AI may eventually get to. It has the potential for major, if not unbounded, levels of targeted enhancement over human capability for advantageous societal gain. On the other hand, it raises parallel concerns about the lack of human understanding of the complexity involved and whether there are sufficient elements of remnant human control to ensure that societal benefits offset any new detrimental threats that may arise. There is, of course, a half-way point where AI is seen as an information adjunct to human intelligence and where the human still has a prominent role in the decision loop.

There is a growing awareness of the need for AI safety methodology research, together with the identification and setting of new safety standards and regulation. The safety standards and regulations already in force for software applications are unlikely to meet the new challenges that AI will inevitably bring. For high-consequence industry applications, this is a paramount requirement; indeed, a research program staying ahead of AI implementation is needed. Therefore, we are likely to see the development of a whole new set of safety standards and requirements of this nature. The current view would suggest that application of AI is unlikely to find its way into warhead design and assembly processes in any extensive way in the near future.

Elon Musk, CEO of Tesla Inc., has noted the dangers arising from industry's economic and survival interests, along with concerns that such research into associated safety may not be keeping pace. *"You have companies that are racing — they kind of have to race — to build AI or they're going to be made uncompetitive; if your competitor*

The contents of this document represent the views of the author and not necessarily those of AWE plc.

is racing towards AI and you don't, they will crush you," he said during a 2017 address to the National Governors Association. This conflict is unlikely to manifest itself with respect to warhead assembly processes where we are not competing in the marketplace.

Technology Revolutions

Some would contend that the human race has seen three major revolutions of advancement technology in benefitting societal aims. These have come with the added responsibility of ensuring that the application of these technologies is undertaken against a set of safety standards, safety assessments and demonstration methodologies.

The First Revolution — The Industrial Revolution enabled us to harness sources of energy to apply to machines, which enabled us to ease the burden of human physical effort in producing wanted products and even opened up the potential for new products. However, from a safety point of view, there was generally human-maintained control over these activities — a human was clearly in the loop and had a prominent role in managing safety.

One other property of this transition was that the activity was generally visual to humans in terms of how processes worked and, as such, there was transparency in the link between intent and action. Although specialist knowledge was necessary to develop the technology, general understanding of how it was implemented could be understood by non-specialists, given a reasonable explanation. In addition, this visual transparency was a clear guide to assessing what could go wrong, together with its consequences and hence the potential impact of the activity with regard to safety. *This revolution was very much about human control, transparency and benefit to society.*

The Second Revolution — The introduction of microelectronics and software with its associated control algorithms was the basis of our second revolution. Here, equipment operation was directly controlled by either hardwired instructions or software set into the machine, but in both instances control was still essentially vested in the human who was responsible for the design of the machine and its internal instructions.

Of course, this transition changed the situation with regard to transparency and complexity. The internal control within the machine and what was happening was no longer explicitly visible, although perhaps the consequences, should something go wrong, could be envisaged. Although the specialist could explain broadly the link between controls and action, it would no longer be transparent (visibly obvious) to the non-expert.

True understanding lay in the mysterious world of embedded hardwired instructions or the even more mysterious embedded software code. The latter took the form of a highly specialized language in the form of complex information sheets together with its microelectronic physical implementation. This took the form of tiny interconnected physical items that gave no visual indication of what was happening inside. This led to a dilemma and the need to ensure that these agents of "mysterious control" did the correct things when required and would not inadvertently do the wrong things, giving rise to concerns for safety, as well as reliability and overall performance. The human requirement of control should not be undermined by lack of complete visibility and understanding of the complexity involved. The "holy grail" took the form of seeking to produce an absolutely complete specification of what the controlling medium should do (and *not* do) under "all envisaged circumstances," and that its implementation was certified to exactly and fully follow the specification.

The benefits from this technology took the form of enhancing automation. It became possible to add sophistication, replacing humans for machine precision and speed, further reducing the human burden and personal risk, and creating a greater range of product capability with consistent quality and cost efficiency arising from mass production. As the sophistication of embedded software increased, so did the need for enhanced understanding of how unwanted outputs (unforeseen) might occur so that they could be detected, eradicated or ensured that they could only occur with acceptable limited probability. This led to two general approaches: looking for a mathematical proof that the implementation perfectly matched a perfect specification, or using a more brute-force approach in which the system was exercised over "all possible circumstances" in a screening process to demonstrate a sufficiently low probability of incurring unwanted outputs. In addition, these were associated with a whole new set of safety standards and related requirements for assessment and demonstration, set to match the level of consequence of failure. This resulted in several universal standards and requirements which were necessary to ensure safe application.

Application of this technology could take two forms: output actions were fully vested in the internal (but human-developed) instructions, or the processed information could be used as an adjunct to guide subsequent human control actions. In the latter case, humans clearly retained final control (tempered with some concern about the probity of the supplied information) over safety. *This revolution was still very much about human control of safety, including the need to develop new safety standards and methodologies to meet the overall goal of benefit to society but with some loss of transparency through complexity.*

The Third Revolution — The next revolution is now potentially with us or "just around the corner," but it is not clear what its ultimate potential might be or how

“Just as the brain can process incoming information and, based on previous experience, use it to make decisions, so ultimately could machines. For example, with the advent of ever-improving sensor and computing capabilities, the question arises: Is it possible to create machines that could match, or even eventually surpass, the capability of humans in terms of perceiving, deciding and taking best-judged courses of action?”

quickly it will be developed and implemented. The second revolution showed the benefits to society that could be accrued from the application of control software and computing in terms of developing new, better and more sophisticated products — generally through its ability to deal in a “routine manner” with complex situations much faster than a human could. Much of this has found its way into automation and robotics.

It has been realized for some time that computing machines can partially mimic the human brain’s “neuron/synapse structure” via the application of so-called artificial neural networks. But while in the second revolution the machine was still subject to human control (human-based design and certification of the controlling software), it is now accepted that further extensions of the brain/sensor analogue in machine/code development can give rise to machine-based learning and a “handing over” of control and decision making to machines themselves.

Currently, AI is primarily based on a learning process that is generally based on human supervision, and the neural network approach is no exception. Although there are other general AI approaches — for example, Decision Trees and Bayesian Networks — it is the neural network approach to AI that is the main subject area of this paper. This approach involves training the machine to recognize objects, conditions, etc. through a process of trial and error while still supervised by humans. This parallels human-based learning where “training and experience” leads to human-informed decision making.

Just as the brain can process incoming information and, based on previous experience, use it to make decisions, so ultimately could machines. For example, with the advent of ever-improving sensor and computing capabilities, the question arises: Is it possible to create machines that could match, or even eventually surpass, the capability of humans in terms of perceiving, deciding and taking best-judged courses of action? By extension, the machine-learning process could even take place under internal machine, rather than human, supervision. This further development of AI, leading to its extension to so-called “Super Artificial Intelligence” (SAI) will bring its own advantages and concerns. For example, will it now become too complicated for full human understanding, and will this raise major concerns over control functions being allocated to the machines? Consider the following:

- “The unpredictability and complex nature of AI presents one of the biggest challenges for humans in understanding its behavior,” said Daniel Kroening, a professor of computer science at Oxford University. “This is why we need to develop AI that will be highly intelligent but transparent enough for humans to understand its complex decisions.” This may become more difficult to achieve as complexity increases.
- There is already practical evidence. “…we can build models, but we don’t know how they work. No one really knows how the most advanced algorithms do what they do. That could be a problem,” wrote Will Knight, the senior editor for artificial intelligence for the *MIT Technology Review*, in his 2017 article, “The Dark Secret at the Heart of AI.”
- A wide range of future concerns were raised at the Beneficial AI 2017 Conference (See Appendix 1).

Difficulty arises mainly through the more-opaque statistical, rather than sequential, deterministic, machine-based decision algorithms. This is good in terms of providing unexpected benefits, but this “incomplete” knowledge is a worry; it could produce detrimental unexpected outputs, or not allow an understanding of *why* safety failures have occurred. Complexity and lack of transparency that gives rise to verification difficulties is of particular concern for high-consequence systems. Coupled with this is the concern that the reasons for failures may be difficult to understand, bottom out and eradicate even after they occur.

This revolution is still underway and there are real emerging concerns over transparency and lack of human control and how to treat its application appropriately.

Intelligence Evolution

Max Tegmark has suggested that the evolution of "what is called life" will fall into three major epochs [Ref.1].

Epoch 1 — This starts at the bacterial level, with little intelligence and some capability to respond to the environment. This progresses through the slow process of the "living" item's evolution to change its form and capabilities in response to environmental influence via natural selection.

Epoch 2 — This is broadly associated with evolution to the human physical form but now where the main and rapid development of capability is associated with wide spread access to technology and information, enabling one to learn and be trained, and to specialize. This is now much more about training the brain, as opposed to further physical evolution of the human form itself.

Epoch 3 — This is broadly associated with the realm of SAI, where machines now surpass the human capability to perceive, decide and act, and where human control plays a smaller and smaller role (at least on the face of things). This presents a vision of a far greater capability to tackle what are seen as today's major unsolved problems affecting humanity. This also begs the question of whether, with the decreasing understanding of how machines work and the gradual loss of human control, machines will continue to decide and act in the interests of human society or, alternatively, in the interests of the machines themselves? For example, the human vision may be the eradication of fundamental medical problems, but this may be of minor interest to machines. Without some overriding form of human control and directive this will not be catered for. By extension, there is the danger that machines might eventually develop their equivalent of "society" — a machine society with all the human characteristics of self-interest, competition, aggression and even direct conflict. Who knows? There is currently no consensus on the limits of AI and what it might eventually lead to in terms of the balance between human societal benefit and risk.

Of course, we are still in the foothills with regard to Epoch 3 with only a limited ability to look into the future and see the possibilities and limits of future AI capabilities, along with any full conception of the associated problems that may arise. With the continuing transfer of more and more decision-making responsibility to machines, new concerns are likely to arise in safety, security and ultimately ethics, coupled with potential concerns about a clash between human and machine interests. These concerns have always been with us in human form in Epoch 2 so in principle there will be no change as such for Epoch 3. Of course, the nature of these concerns will be markedly different.

Innovation

One may ask whether AI can match the human attribute of *innovation*. Do human ideas turn up like the turning on of the proverbial "light bulb," without any idea of where they come from? Or do they arise from the multitude of past experiences stored in the brain, which are formerly somewhat unrelated but subsequently are brought together in more cohesive fashion at some point in time to give rise to the new ideas — innovation? If it's the latter, in principle, there appears to be no reason in principle why AI machines cannot progress to the state we humans call innovation, but Roger Penrose disagrees [Ref. 2]. The parallel machine process of learning, or even self-learning, and trial and error will generate such latent experiences and stored knowledge that it can give rise to the germinating of new ideas or the equivalent of innovation. Is there a machine analogue of the "light bulb" effect?

Currently AI machines are more directed toward answering questions already set by humans. Machines with the ability to decide for themselves which questions need to be posed and then answered, looks like yet a further step in AI evolution.

How Far Will AI Go? Software Aspects

Some suggest that there is nothing in the laws of physics that identifies *any* limits to the development of SAI. The brain is a vast collection of neurons and synapses, with stored information and linked pathways influenced by past experience (memory) sensed through our range of sensor systems and modified (trained) by success or otherwise in correct decision making.

In short, at any time the brain's operation is fundamentally based on the laws of physics (and chemistry) given what exists. This process is mirrored in AI through the application of neural networks or so-called "deep learning," similarly governed by the array of storage elements and linkages modified by "past experience." This past "experience" gained through the seeing, testing and learning processes (memory) gives rise to the associated algorithms that enable the decision-making process. Such algorithms can aid in the recognition of targeted features, detect other unsuspected features and aid in finding the optimum path to reaching a given requirement. AI success or otherwise comes from the level of correctness and completeness of the training information that acts as a moderating influence on the neural network approach. The analogy of human judgement and common sense? As such, there appears to be no limit to the sophistication and complexity associated with computation and AI decision algorithms, given a suitably sophisticated hardware platform. However, there are limits to the level of problem complexity that so called "classical" computers can handle.

Eventually, such complexity will push machines into conditions where quantum interference arises that may limit further capability. In fact, this takes us into the next dimension, with so-called quantum computers that are based on quantum principles and associated quantum-based algorithms. These offer even greater capabilities for handling a wide range of problems and specifically enhanced AI. Of course, once machines have reached a superintelligence standard, it is difficult to see how such learning and enhanced AI algorithms development can continue to be based on human intelligence intervention. In fact, such learning will need to be machine self-initiated rather than through human intervention because humans may no longer have the level of intellect to develop artificial intelligence further. Will there be no end to this?

> “We are now approaching the stage where quantum computers offer a completely new computing regime and where early and somewhat limited devices based on these concepts have been successfully demonstrated. The usual classical computer deterministic binary element of either ‘0’ or ‘1’ is replaced by the probabilistic quantum bit or qubit which can simultaneously have values of *both* ‘0’ and ‘1’ through the quantum principle of superposition.”

How Far Will AI Go? Hardware Aspects

There has been a steady and dramatic enhancement of sensor and computing technologies over the last few decades with Moore's Law broadly describing how computing speeds and capability increase as the computing elements — and expense — get smaller and smaller, and with greater efficiency seen in the interconnection and parallel processes. Will these processes reach a limit when the elements involved in sensing, storage and computing operations themselves approach some physical limit? Currently, "switching" elements have reached the "few atoms" scale and memories are possible even at the subatomic particle regime. For example, there are already memory elements based on "particles" with binary spin states. In fact, switching elements are now reaching the stage where quantum effects such as tunneling are beginning to show their influence and potentially limit further enhancement of capability.

We are now approaching the stage where quantum computers offer a completely new computing regime and where early and somewhat limited devices based on these concepts have been successfully demonstrated. The usual classical computer deterministic binary element of either "0" or "1" is replaced by the probabilistic quantum bit or qubit which can simultaneously have values of *both* "0" and "1" through the quantum principle of superposition. These qubits can, in turn, be processed through processes of entanglement, analogous to the mathematical operations in classical computers, to produce targeted results. The errors are bounded by the ability to maintain coherence during the overall process. Such quantum computers will have the ability to tackle problems that are impossible with classical computers where the exponential demand on their capability arises as complexity of the task increases. The superposition nature of the qubit in the quantum computer is the key to dealing with this complexity, together with the entanglement processes which effectively enable many operations to be undertaken simultaneously as opposed to the linear approach in classical computers.

To paraphrase Richard Feynman, *"If you want to solve a problem of quantum level of complexity you need a quantum computer."* In fact, there is already evidence that researchers are using quantum computers in machine-learning mode in an endeavor to solve quantum problems that have not been successfully solved by other means.

The bottom line is that these physical enhancements may well show more rapid development than the further physical evolution of the human brain itself. In turn, the associated learning processes will influence how these machines further develop themselves in the hardware context — that is, the machine will take over the responsibility for better and more efficient machine design and its implementation. Are there fundamental limits for these processes?

AI and the Role of Specialists

Of course, the old adage of "garbage in/garbage out" still very much applies to AI. The role of the specialist remains crucial in ensuring the correct neural network structure and that the appropriate quantity, fidelity and balanced set of training and test data are used together with independent review of transparency of the process that leads to the machine's results. In addition, success will also be based on the probity of the associated algorithms. There are many examples where follow-up human-based forensic assessment has shown flaws in all of these aspects. When the machine conclusion defies common sense, such errors flag an obvious need for re-assessment, but more subtle errors can still lead to misleading results.

> “Many noted thinkers, including Stephen Hawking, Elon Musk, Bill Gates and Eliezer Yudkowsky amongst others, have raised concerns about super intelligence — particularly that machines might reach the stage of looking after their own interests and destiny, rather than those of society and the humans they are ‘designed to serve.’”

As complexity increases, so will the difficulties with regard to executing this human intervention. The output from AI execution will not explicitly identify the physical basis on how the machine's conclusion was based. It will be bound up in complex statistical analyses with associated "confidence levels." It is also currently unclear whether there will be tractable processes by which the analysis can be viewed in the "reverse direction" to pinpoint the core aspects supporting any (wrong) conclusion. Highly skilled humans will still be necessary to seek answers to these problems

The experienced human specialist will always be important as an adjunct to the machine (or the machine as an adjunct to the human) as a player in the overall information monitoring, interpretation and decision loop.

A Vision of the Future — Impact on Humanity's Interests

One may visualize a time when humans no longer have the level of intellect to provide the learning process for machines to further enhance AI. In addition, the time may come when humans will no longer be capable of fully understanding how these machines operate and the processes by which they make decisions. Is this the "sea change" when machines become independent and self-sufficient, taking the role of self-learning, self-testing and self-designing, for the purpose of continued intelligence enhancement — *and self-interest?*

- *"What we really need to do is to make sure that life continues into the future .... It's best to try to prevent a negative circumstance from occurring than wait for it to occur and then to be reactive."* Elon Musk [1].

Many noted thinkers, including Stephen Hawking, Elon Musk, Bill Gates and Eliezer Yudkowsky amongst others, have raised concerns about super intelligence — particularly that machines might reach the stage of looking after their own interests and destiny, rather than those of society and the humans they are "designed to serve." These interests may well be in conflict with each other and this could have a detrimental impact on society.

There is no consensus on how artificial intelligence will progress with time, when it will reach the level of human intelligence and when it will reach the stage of solving some of the fundamental problems that are currently beyond human capability. Some perceive there is no fundamental limit to AI's capabilites and, in turn, this enhanced intelligence may uncover as yet undiscovered major problems that themselves will need urgent solution. As noted previously, this potential for enhanced human benefit comes with a fresh set of associated concerns — particularly if machine independence turns into machine self-interest that conflicts with what is best for society.

- *"If a machine can think, it might think more intelligently than we do ... This new danger ... is certainly something which can give us anxiety."* Alan Turing [2].

Turing suggests *"turning off the power at strategic moments"* as a possible solution to discovering that a machine is misaligned with true human objectives. However, a super-intelligent machine acting in its own self-interest is likely to have taken steps to prevent this!

- *"If you're not concerned about AI safety, you should be. Vastly more risk than North Korea."* Elon Musk [3].

[1] Said in 2015 after donating $10 million to the Future of Life Institute (FLI) for research into keeping AI safe.

[2] Said during a 1951 lecture on BBC Radio 3

[3] August 11, 2017. 8:29 p.m. Tweet.

The Impact on Safety Standards and Assessment Methodologies

Currently, there are two distinctive approaches to software-related safety:

Decision Making that is Essentially Human Controlled — By and large, this regime is based on deterministic sequential *cause and effect*-based hardwired/software-based processes, with Monte Carlo-type analyses often implemented when there are uncertainties (for example, in weather prediction). The internal process will take the form of instructions that are as complete as possible, identifying what should happen and what should not happen, given any external input to the system.

The overall process is based on a "requirements specification" that attempts to cater to all envisaged possibilities. Final evaluation for completeness in relation to the specification — and its faithful representation in the hardwired or coded implementation — takes the form of extensive "cold" exercising of the system for certification before application to the intended purpose. This, in principle, gives a "practical" assurance measure of the accuracy and completeness of the conversion of the specification into machine language (limited, of course, by the completeness and accuracy of the specification itself). The required level of confidence will be bound up with the level of consequence of getting it wrong.

This overall balanced approach generally follows the Safety Integrity Levels (SIL) process, which sets down a rigorous, evidence-based, structured and transparent confidence/risk approach with a detailed audit trail (see IEC 62061, for example). A SIL associated with a consequence/frequency matrix is generally described by a discrete level for specifying appropriate integrity requirements for safety functions to be allocated to safety-related systems. SILs are typically assigned to four levels, where Level 4 is the most stringent and Level 1 the lowest.

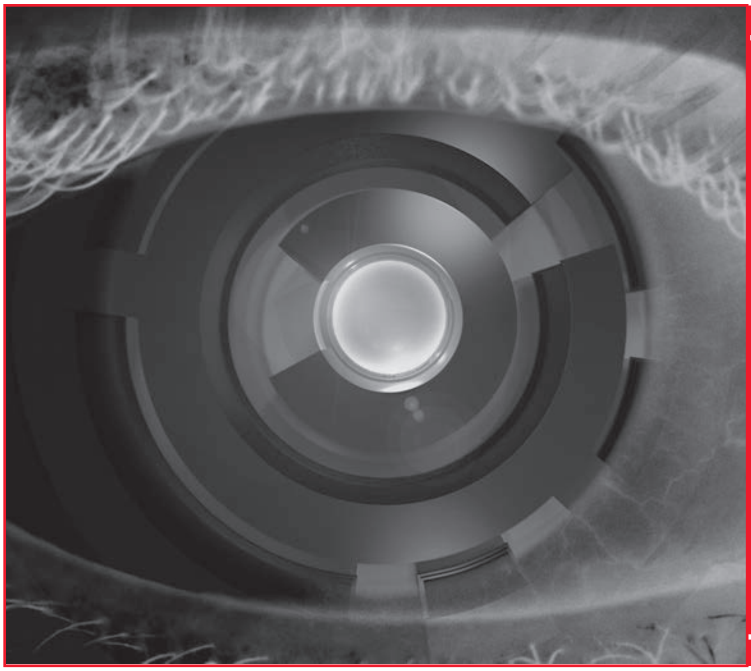
Decision Making that is Machine Controlled — With the increasing advent of large data bundles and the added sophistication of computing machines and related software there is an increasing trend towards decision making based on machine intelligence. This will give rise to a need for revised safety standards, certification, regulation and associated assessment methodologies. Many recent AI advancements have taken place in the context of artificial neural networks, which somewhat mimic the human brain in a parallel processing sense, but to a much lower level of overall capability in the general sense. This trend has been enhanced by developments in parallel computing and aided by advanced graphic processing chips (GPU), where the latter packs thousands of relatively simple processing cores (the effective neurons) on a single chip. The neural net will consist of an input set of neurons, an output set and intervening hidden layers. For optimum precision and efficiency, the network structure and its interconnections are chosen to best match the specific case under consideration — as are the algorithms that control the search for the optimum weight conditions that produce the best matches of its outputs to expected outputs during the machine's training phase.

Machine learning and related decision making are almost exclusively bound up with probability functions and associated confidence levels as opposed to deterministic processes. These statistical assessments are generated following the teaching and testing phases. The machine is judged to be trained and tested when it consistently provides the correct outputs for a wide range of input data containing examples of such outputs. The network accomplishes this essentially during the training phase where its internal weights are continually adjusted until there is minimal error between machine assessment and expected outputs. The network is trained in this manner to identify all expected input conditions.

Of course, the overall process is far more sophisticated than what appears to be a "trial and error" approach. The optimizing process takes place through a complicated stepped process for minimizing the error. This is typically undertaken by a reverse error propagation process called "stochastic gradient descent," where the gradient of the path leading to the expected result is evaluated for each weighting factor. This leads to a global error minimum. An activation parameter is chosen to provide continuous functions and continuity in the gradient evaluation approach. A simple case is represented by a machine whose job is simply to identify a specific pattern that should then trigger a response action, which could be a safety-critical action. For example, in a flight control application, the machine may detect what it recognizes as the impending unsafe failure of a flight-related component and identify the need to take a safety mitigating action. On the other hand, there is the opposite concern of a false positive causing a mitigating action which might in itself lead to danger. Generally, machines may be required to identify several separate patterns with a specific control response allocated for each. In this case, statistical confidence in discriminating between these patterns should not be unduly impaired (decision error) by the overlap of the statistical probability tails.

Should AI be Used? Risks and Benefits

The decision of whether to apply AI will depend on many factors that weigh potential *benefits* against *risks*. In fact, the best strategy often takes the form of using the best attributes of both machine and human input in an optimized approach. In addition, more than one independent approach to AI application can be ap-



> “The decision of whether to apply AI will depend on many factors that weigh potential *benefits* against *risks*. In fact, the best strategy often takes the form of using the best attributes of both machine and human input in an optimized approach. In addition, more than one independent approach to AI application can be applied for redundancy purposes. Decisions will be partly governed by the level of consequence arising from errors, as well as the uncertainty as to why errors have occurred and how they may be rectified. After all, AI is a statistical approximate rather than a deterministic approach.”

plied for redundancy purposes. Decisions will be partly governed by the level of consequence arising from errors, as well as the uncertainty as to why errors have occurred and how they may be rectified. After all, AI is a statistical approximate rather than a deterministic approach. This does impact heavily on any risk/benefit decision about application for high-consequence industries. Nevertheless, such applications are already taking place in a number of high-risk industries. For example, aviation flight control (not without recent concerns), chemical engineering, power plants, automotive control and medicine among others — all where humans still have a prominent role in the decision process. However, there are other examples such as the design, manufacture and ownership of nuclear weapons where there will always be a strong resistance to the transfer of human control to machines either for automation or AI purposes.

Items which Influence the Machine Decision-Making Approach

Influences in the machine decision-making approach can be broken down into several areas. These include:

Validation — Here, validation is used in the sense of making the benefit/detriment analysis (including the detriment of *not* applying AI) together with selecting the best option that includes the “As Low As Reasonably Practical” (ALARP) process for safety. Validation includes the best application of combined machine and human contributions in the overall strategy, or even includes the application of more than one independent AI approach. Benefit comes in many forms: greater efficiency, greater accuracy in performance, quicker response, less cost, coping with complexity, mass production, etc. On the other hand, this may be offset by the level of consequence of a machine getting it wrong, or the cost of the machine overhead given its limited level of application. The approach will be very much dictated by the application. For example, the nature of the risk/benefit balance will be different when producing chocolate bars than it is when producing nuclear warheads.

The following ideas should be kept in mind when considering this approach:

1. Neural networks are universal approximators, and they work best if the system application has a high tolerance to error.
2. All AI-based application and decision making in isolation should be used with great caution where failure can lead to catastrophic consequences.
3. There should be a compelling evidence-based assessment of the benefit advantage over associated detriments before relying too heavily on AI-based decision making.
4. There should be a well-established research and development base to support any decision that relies heavily on AI-based decision making.
5. Decisions to apply AI should not be overly driven by the needs of the competitive marketplace.

Validation and Verification — In this case, *validation* means an assessment of whether the overall approach is best matched to purpose (“Doing the right job!”). *Verification* means that the approach is supported by the appropriate level of evidence that quality standards have been met (“Doing the job right!”). The discrimination capability and efficiency of a network will be associated with its structure; that is, its number of neurons, layers and interconnections and the associated training algorithms. All will have an impact on the "residual discrepancy value" (accuracy) during the training process and the confidence in the output judgement.

As we enhance such structures in what is termed "deep learning," we inevitably encounter more complexity. This, in turn, can give rise to further loss of transparency in how the neural network actually does its work in a cause-and-effect sense.

Key ideas for V&V include:

1. The neural network configuration should be optimized (and not over-built) to match the application. This aids in the efficiency and the statistical accuracy arising from the training and testing procedure.
2. Software algorithms associated with the training programs should go through an appropriate V&V process for correctness in application, and should be vetted to eliminate unintended biases. There are now open-source algorithms of this nature available.
3. The statistical-based software involved in assessing confidence levels should also go through a V&V process.
4. The training program includes training the machine to correctly recognize patterns. By illustration, various versions of dogs are presented as input and the machine is trained to recognize these as dogs. The training sets and program should be comprehensive, complete and error free, as well as sufficient to uniquely identify and cover the expected outputs with sufficient confidence. The size of the structure will depend on the number of expected outputs, the level of discrimination and the required statistical confidence levels. Of course, these can include those outputs that are sentenced as fault or failure conditions with required mitigation actions.
   The requirements and residual questions may include:
   a. Training data should be compatible (easily transcribed) with the machine input interface.
   b. What are the standards for completeness, sufficiency and accuracy, balance and lack of inadvertent bias necessary in the training data and program? This is an area which can be prone to subjectivity and inadvertent bias. Can generic standards be developed, or will they always be strongly driven by specific application?
   c. Should cover all envisaged input conditions expected in the final applications.
   d. Should avoid potential inadvertent (or subjective) bias
      i. Subject to independent scrutiny
      ii. Bias and error can arise via the choice of training sets and by erroneously labelling a result as correct, e.g., telling the machine that an input example represents a dog when, in fact, it doesn't.
   e. Have all possible expected outputs been catalogued, and do they form part of the training program?
   f. How should the statistical confidence level be set for expected output detection and discrimination between expected outputs influenced by "failure consequence"?
5. The testing program, on the other hand, subjects the machine to a separate set of data (usually a subset of the training data that is not used in the training process). The machine is now blind-tested on its ability to discriminate correctly between patterns. If the result proves to be unsatisfactory, the machine should be subject to further evaluation, training and testing.
   Again, the requirements and residual questions would appear to be:
   a. That the form of the test data should be compatible (easily transcribed) with the machine input interface.
   b. What are the standards for completeness, sufficiency, accuracy balance and lack of inadvertent bias necessary in the testing data and program. Can generic standards be developed, or will they be strongly driven by specific application?
   c. Should cover all envisaged input conditions expected in the final applications.
   d. The test data should lie within the envelope of the training data; otherwise it may make erroneous decisions
   e. What should be the resultant level of identification confidence — weighted by the level of error consequence?
6. The final stage will require a monitoring activity for AI performance when applied to the targeted process to gain further assurance or identification of unexpected events and rectify them in a continuous process of Review Learn and Improve (RLI).
   a. The application should lie within the envelope of the training and test data
   b. Continued monitoring, recording and sentencing of output anomalies should include
      i. Level of monitoring
      ii. Level of anomaly discrimination
      iii. Causes of anomalies
      iv. Rectification
      v. Identification of weakness in the structures, training, test data and algorithms
      vi. Upgrade of the structure, test and training cycle

7. Are machines misled by false association? Do the internal processes in the machine themselves produce false biases? For example, the machine may associate a correct result with a pattern that often appears when correct outputs are given during the training process. However, the appearance of this pattern does not necessarily imply that the correct criteria have been satisfied and the machine can hence be misled. For example, a dog might be often associated with a specific background pattern. The appearance of such a pattern should not imply a dog.
8. There may be subsequent changes in the system of which the AI forms part or even modifications of purpose. This may necessitate further elements of training, testing and neural network (including algorithm) reconstruction and even reassessment of the best strategy for machine and human contributions in the overall decision-making process.
9. As noted previously, confidence can be augmented by redundancy, using independent (in technology, training and testing) AI machines for the purpose of averaging, range of disagreement, or voting strategies and minimizing inadvertent false bias.

> “The neural nets undertake their internal processes ‘without’ external transparency and ‘appear’ to come up with the required results with accepted levels of being ‘error free.’ There is often no longer a clear indication of a relationship between *cause and effect*. Does the apparent picture of ‘it is working well’ give a guarantee that it is free from unexpected errors? There is a nagging concern about products and processes where such errors can lead to catastrophic results (the potential for a ‘Black Swan’).”

Integrity — The well-established SIL process, which sets down a rigorous, evidence-based, structured and transparent confidence/risk approach with a detailed audit trail, appears to provide a suitable framework for the AI approach. This would be undertaken in association with a complementary ALARP case to show that the best approach had been undertaken and that any further effort to reduce risk further would be disproportionate to the reduction. The only difference lies in the specific contributing elements that are uniquely associated with the AI approach and listed above. *This is where most of the hard work, assessments and judgments need to be carried out in order to provide the evidential basis in respect for the required SIL level.* The level of risk will be set by the details of the application, and the integrity level requirement will be based on a customized risk/SIL matrix structure.

The Black Box — Unlike the more traditional sequential, deterministic software application, where the human is effectively in control via specification through to software implementation, there is a less direct hands-on human control with machine learning — particularly for the case of neural networks. The neural nets undertake their internal processes “without” external transparency and “appear” to come up with the required results with accepted levels of being “error free.” There is often no longer a clear indication of a relationship between *cause and effect*. Does the apparent picture of “it is working well” give a guarantee that it is free from unexpected errors? There is a nagging concern about products and processes where such errors can lead to catastrophic results (the potential for a “Black Swan”). Therefore, AI application to critical parts of nuclear warhead design and manufacturing without human elements of control must be viewed with extreme caution. An understanding of how the AI process leads to “success” is currently one of the main and most important avenues of AI research in neural networks. This understanding aids in enhancing confidence and points to those areas where uncertainty remains and where further work is necessary.

Approaches to error minimization are being made through detailed internal tracking of what is happening in the neural net as it progresses through the steps to error minimization. The process does appear to home in on those elements of the input pattern that it regards as core to the recognition process and, in turn, gradually discards those which it deems as extraneous as it runs through its training process. For example, pictures of different dogs in different settings appear to be quite different. However, the process seems to home in on those elements which it regards as key for recognition of “dog,” rather than those associated with the general setting — despite that, there may be some relationship between a setting and a dog being present. This would become an unacceptable bias if used in the network approach (machine misled by false association). However, for some applications associations of this nature may be seen as an aid to classification.

1. Any machine certification program would need at least some supporting qualitative evidence of understanding of the “machine’s logic” to counter any concerns raised by the Black Box nature of the process.

2. Success in this avenue of research would obviously be key to the application of AI to high-consequence industries.

Safety and Security

Just as AI can lead to enhanced safety if properly implemented, it can also lead to enhanced security. For example, AI can be trained to identify patterns that indicate potentially unsafe conditions and then take mitigating actions; it can also be trained to detect patterns that indicate potential security anomalies and act upon them. Safety concerns arise from natural or inadvertent causes and security anomalies arise from both malicious and accidental actions. In each case, the potential output, if not detected and mitigated, produces a detriment. However, in both cases the process will be similar in terms of training and testing the AI to detect patterns that identify unsafe or unsecure conditions with the need to develop mitigating procedures. Just as in the safety case, the level of success will be based on the quality of meeting the requirements identified above, but where the training and testing and monitoring will be based on a set of envisaged security anomalies.

Conclusions

We are now very much entering a new regime with AI, where there is a gradual trend of handing over decision making from humans to machines, and where the application of neural networks plays a greater and greater role. Although this enables one to provide products and services that offer significant advantages, these developments also carry with them new safety concerns. The danger is that implementation may be excessively driven by the needs of the competitive marketplace and that parallel research into establishing safety standards, regulation and methodologies may not keep pace with the rate of application.

The purpose of this paper has not been to define new standards, but to point out in a general manner where such standards need to be established and applied. It is generally agreed that neural networks are universal approximators, and they work best if the system application has a high tolerance for error. The overall criterion should be: do the advantages of such application outweigh the detriments and, in particular, the level of risk incurred? This is by no means a trivial assessment for an organization which deals with high-consequence product design and processes. Although no organization can stand still in the presence of such developments, acceptance of AI's application must be based on a solid foundation of safety-based research, standards, certification, regulation and understanding.

AI represents a paradigm shift in capability and in ways of working for organizations. In turn, this also presents a paradigm shift for those organizations that have an independent responsibility for setting standards and executing regulation. Concern still resides in the sense that it may not be possible for some time to fully understand how neural networks actually reach their decisions and whether this precludes unfortunate surprises. This represents a major concern for organizations where such errors in their product design and processes can lead to catastrophic results and as such there will be no rush to exclude humans from the decision loop.

Appendix 1: The ASILOMAR Principles

The following 23 principles, set under three broad headings, were advocated following the Beneficial AI 2017 Conference and have been generally accepted by a large number of people who work with or have an opinion on Artificial Intelligence. They are mainly directed towards protecting against the potential negative concerns that may arise as AI continues to advance. They are far reaching both in terms of scope of coverage and look far into the future. Those that directly relate to safety are in bold text.

Research Issues

1. Research Goal: The goal of AI research should be to create not undirected intelligence but beneficial intelligence.
2. Research Funding: Investment in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics and social studies such as:
   a. How can we make future AI systems highly robust so that they do what we want without malfunctioning or getting hacked?
   b. How can we grow our prosperity through automation while maintaining people's resources and purpose?
   c. How can we update our legal systems to be more fair and efficient, to keep pace with AI and to manage the risks associated with AI?
   d. What set of values should AI be aligned with, and what legal and ethical status should it have?
3. **Science-Policy Link: There should be a constructive and healthy exchange between AI researchers and policy-makers.**
4. **Research Culture: A culture of cooperation, trust and transparency should be fostered among researchers and developers of AI.**

5. **Race Avoidance: Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.**

Ethics and Values

6. **Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.**
7. **Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.**
8. Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by competent human authority.
9. Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse and actions with a responsibility and opportunity to shape those implications.
10. Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
11. Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms and cultural diversity.
12. Personal Privacy: People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.
13. Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
14. Share and Benefit: AI technologies should benefit and empower as many people as possible.
15. Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity
16. **Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.**
17. Non-Subversion: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert the social and civic processes on which the health of society depends.
18. AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

Longer Term

19. Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capability.
20. **Importance: Advanced AI could represent a profound change in the history of life on earth and should be planned for and managed with commensurate care and resources.**
21. **Risks: Risks posed by AI systems, especially catastrophic or existential risks [cause human extinction or permanently and drastically curtail humanity's potential], must be subject to planning and mitigation effects commensurate with their expected impact.**
22. **Recursive Self-Improvement: AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.**
23. Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals and for the benefit of all humanity rather than any state or organization.

About the Author

Malcolm Jones previously led the Distinguished Scientists group at the Atomic Weapons Establishment (AWE). He currently holds the position of Scientific Adviser to AWE's Chief Scientist and directly supports AWE's Chief of Product Assurance. His career at AWE has taken him through a wide range of scientific and engineering topics, but he has maintained a continuous association with nuclear weapon design and process safety and top-level nuclear safety standards. His interests extend to corporate safety cultures and the root cause reasons for failures. He is a Fellow of the International System Safety Society and is an adviser to a number of senior U.K. Ministry of Defence and AWE safety bodies. He has been awarded an MBE in the Queen's Birthday Honours List for contributions to the U.K. defence industry and is a recipient of the John Challens' Medal, which is AWE's highest award for lifetime contributions to science, engineering and technology. He has also been honored by VNIIA in the Russian Federation for his work in fostering nuclear weapon safety collaboration between the U.K. and the R.F. ⬢

References

1. Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence*, Knopf, New York, 2018.
2. Penrose, Roger. *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*, Oxford University Press, Oxford, U.K., 2002.