Artificial Intelligence: An Alternative View

KEEL® Technology (Explainable AI)

Auditable and Explainable Adaptive Intelligence

Compsim White Paper

There is a recent upsurge in press coverage on the topic of Artificial Intelligence. The US Government has initiated an interagency working group to learn more about the benefits and risks of AI.[1] In the commercial space, some organizations are attempting to redefine the market area based on a specific approach. Perhaps it would be appropriate to revisit the original concept of artificial intelligence and identify services that are really desired, and then compare them to some potential outcomes that may not be so desirable.

If you start looking for a definition of Artificial Intelligence, you are directed toward the "artificial" aspects. So perhaps it is more appropriate to look at a definition of Intelligence first:

> **"Intelligence** has been defined in many different ways including one's capacity for logic, understanding, self-awareness, learning, emotional knowledge, planning, creativity and problem solving. It can be more generally described as the ability to perceive information, and retain it as knowledge to be applied towards adaptive behaviors within an environment or context.**"[2]**

Note the objective of providing "Adaptive Behaviors". Now expanding the definition to Artificial Intelligence:

> **"Artificial intelligence (AI)** is intelligence exhibited by machines. In computer science, an ideal "intelligent" machine is a flexible rational agent that perceives its environment and takes actions that maximize its chance of success at some goal. "[3]

Perhaps a more general definition might be:

> **Artificial intelligence** is ***machines delivering expert adaptive behavior***.

In this definition there is no reference as to how the intelligence (expertise) is created or provided. It is simply a qualified noun. A user of artificial intelligence will use it because it offers value: better / faster decisions, and / or better operational control. There appears to be

[1] The White House Blog: Preparing for the Future of Artificial Intelligence; https://www.whitehouse.gov/blog/2016/05/03/preparing-future-artificial-intelligence

[2] Wikipedia: https://en.wikipedia.org/wiki/Intelligence

[3] Wikipedia: https://en.wikipedia.org/wiki/Artificial_intelligence

some industry effort to redefine artificial intelligence to incorporate how knowledge or expertise is acquired into the definition, by suggesting that an artificial intelligence must be able to learn.

Perhaps there is value in looking at two different objectives of artificial intelligence and consider the where we are in the evolution of AI.

Two Different Objectives of AI:

1. Deliver expertise in machines: decision-making and/or expert operational control.
2. Use machines to develop expertise: create expertise where knowledge and understanding does not exist.

By separating artificial intelligence into these two application areas, it may be helpful to refine one's objectives.

Delivering Expertise in Machines:

If the objective is to deliver expertise in machines, then the focus is on how to interpret and react to changing information in a changing or diverse environment (delivering adaptive expert behavior).

One might suggest that when you have machines delivering expertise, an equal or more important service will require the machine to explain (with mathematical precision) its decisions and/or actions.

> "When you can measure what you are speaking about, and express it in numbers, you know something about it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarely, in your thoughts advanced to the stage of science."[4]

Perhaps providing some examples of Machines Delivering Expertise would be helpful in understanding why the explanation of decisions and actions is important.

Self-driving cars

If you are going to trust your life to a self-driving car and the expertise of the cognitive machine controlling it, then you should expect it to be able to explain its behavior… especially when the vehicle encounters a situation where there are no perfect outcomes such as a scenario where

[4] William Thomson, 1st Baron Kelvin; http://www.goodreads.com/quotes/166961-when-you-can-measure-what-you-are-speaking-about-and

someone is going to get killed or injured. The vehicle will be balancing risk and reward to get you to your destination. You probably want to know how the car values your life. Assuming your self-driving car has some assigned or derived self-value, and assuming that other self-driving cars have their own self-values (for example, their assigned or derived ethics, and their assigned or derived expected operational characteristics), these vehicles will all be balancing the risks and rewards in transporting you from place to place. Sometimes unexpected things will happen: contaminated sensors, unexpected environmental situations, or possibly a random human doing something stupid. The self-driving car may make a decision that results in the loss of life. You cannot hold the vehicle responsible (unless the machine learned how to drive completely on its own, and you accepted that fact when you got into the vehicle). One might hope that the self-driving car supplier would also want to perform an after-accident review to insure that the operational policy (how it interpreted the information) was correct. And, if there was a problem with the policy, it could be quickly identified and fixed after the review.

Adaptive medical treatment

The Internet of Things (IoT) provides for connected intelligent sensors. There is nothing prohibiting the adaptive treatment of disease: from adaptive diagnoses through adaptive treatment. As these systems are introduced to continuously measure the body's response to treatment and compensate for individual personalized responses to medication, one would hope that there is some oversight to that treatment, especially if things go wrong. In the area of personalized medicine a single drug can cure one person and kill another, so adaptive dispensing will be critical. Correcting what doesn't work will be mandatory as this technology evolves. If you cannot measure it, you cannot fix it.

Autonomous Combat Weapon Systems

The use of fully autonomous weapon systems will happen, whether we are ready for it or not. Humans are just too slow to be involved in all of the decisions. Plus humans make mistakes, just because they are human: lack of attention, failure to recognize a situation, overwhelmed with information, poor judgment. It is generally recognized that you cannot trust a weapon system that learns completely on its own. It might decide to act just like some humans that switch sides during a conflict. This means that human driven policies should control the behavior of the machines. It also means that after-mission reviews need to be performed to validate the behavior of the systems. Currently when after-mission reviews are held for human controlled systems, the review primarily focuses on the behavior of the human. When operational policies are given to machines in mathematically explicit ways, the reviews will focus on the quality of the policies and the information used by the machine while executing those policies. It will be (one would hope) mandatory that the policies and the information that drives them can be easily understood by humans reviewing the activities of the system. As this area evolves, new sensors will be requested and developed so more appropriate decisions and actions can be delivered. The bottom line is that if you cannot review the information fusion model of the machine, you cannot correct and enhance the systems. When an organization releases autonomous combat weapon systems into the field, you cannot hold the machine responsible. Rather, it is those fielding those weapons that will be held responsible. So it is in everyone’s best interest that those machines and their operational policies can be studied and fixed if necessary. Remember that these machines will be mass produced, so the potential for massive error is amplified.

Battlespace Management

Battlespace management (like corporate strategic management, or even management of professional sports teams) uses written or many times unwritten policies to guide the big picture decisions and actions. In many human driven systems this is a gut feel process. But as the drive to do more with less, and the cost of less than optimal decisions rises, there will be a demand to automate more of the “business processes”. While some might look at battlespace management as a domain where point decisions are made, this area will evolve into a domain where one is constantly evaluating the battlespace searching for momentary weaknesses and for opportunities to arise. This means the controlling systems will be constantly looking for new opportunities and new threats, and constantly adapting to the situation. “Gut feel” policies will need to be translated into mathematically explicit policies that can be executed by machines. At first this will be used to audit human decision-makers and warn them of decisions or actions that might present a significant risk. In the future this will evolve to an advisory role where decisions and actions will be suggested to the decision-makers. It will always be important that the decisions and actions suggested by the machines can be easily audited and explained. In these kinds of operational policies, the opposition always has a vote. Their tactics and strategies will change. If you cannot explain why / why not / how / how much / when / and where, your audience will (hopefully) be reluctant to trust the decisions and actions suggested by the machine. And, if the decisions and actions cannot withstand the review, they should be rapidly corrected and augmented with additional information sources.

Use Machines to Develop Expertise:

There appears to be a commercial attempt to integrate how knowledge and expertise is captured into the solution delivery process. Perhaps this is not the best approach. If the knowledge creation process is tightly coupled with the delivery of the acquired expertise a less than optimal solution may be provided. This may be because the collected understanding may not have been given the appropriate level of scrutiny. The objective is normally just to deliver expert adaptive behavior and you need to trust and review that behavior.

This questionable linkage between machine learning and artificial intelligence might be restated as acquiring knowledge versus applying knowledge. One might suggest that machine learning deals primarily with the accumulation of data rather than the delivery of expertise. This doesn’t mean that the automatic development of knowledge and understanding should not be pursued. It just suggests that the development of knowledge and understanding may have a different timeline than the need to deploy expert behavior.

One might also suggest that at least some consumers of artificial intelligence may be very interested in obtaining a detailed explanation of any decision or action delivered by the AI system, especially when there may be both positive and negative consequences of a decision or action. So the following questions might be asked of a machine learning system:

- Is there real value in collected intelligence that cannot be explained?
- Is it possible to trick machine collected intelligence?
- Have humans had a hand in machine collected intelligence that might bias the results?
- If machines collect intelligence, how sure are you of the results? (Does it matter?)
- If machines collect intelligence solely on their own, do you know how the machines might be biased?
- Can the machine explain decisions and actions derived from learned knowledge?
- Who is responsible for machine derived intelligence?

Machine Learning versus Human Learning

"*Trust me!*" If a machine said "*Trust me, not those stupid humans!*" how would it be perceived? Or "*Trust me, the world is flat. Humans are stupid. Trust me! I am a computer, and what I say is true.*" Or, "*Trust me. Humans are not responsible for climate change. Trust me. Burn coal. Burn, Burn, Burn.*" If a machine learning engine was given access only to the Flat Earth Society publications and publications from the coal industry lobby, do you think you might get these answers from the machine? Who is responsible? Is the machine responsible?

It is a well-known marketing tactic to use repetition to create a belief. Every time a statement is repeated, there is some subliminal bias established in the human brain. Marketers know it. Politicians know it. Can a computer publish opinions very fast in a form that another computer might read and interpret? Can this set a bias in a machine learning system? Can an artificial truth be created?

It may be obvious that this is a battle for the control of the human mind: a desire to control human behavior. Organizations have been attempting to find a way to do this for years with brain washing techniques, drugs, deep brain stimulation, and techniques to turn on / off parts of the brain. One can easily translate these objectives to be a battle to control how machines think and act. They will do this by controlling how they think as well as where they get their knowledge and information.

Restated, this might be relabeled as a battle for control over the mind of the machine. If a machine learns from "big data", then if you (or someone, or something) controls "big data" then they/it can bias the machine's decisions and actions.

Some humans are skeptical about the objectives of other humans. This is an inherited protection mechanism that helps protect the species from extinction. It remains to be seen if the same humans are *as skeptical* of decisions and actions suggested by a machine. Some (not so technical humans) may believe that if decisions or actions come from a computer, they must be correct. One would hope they are just as skeptical, or even more skeptical, because the machines can be mass produced. If those machines are inappropriately biased, then (hopefully) they will be continuously audited and the policies will be adjusted accordingly. The potential impact of inappropriate decisions and actions is amplified because the machines delivering decisions and actions based on that information can be mass produced. If the machine cannot easily explain its decisions and actions to humans, then one might suggest that humanity will be at risk.

This section focuses on the present state-of-the-art in machine learning. Because humans are involved in selecting the information available to the machine with the purpose of gathering knowledge, one might suggest this is more "information processing" than it is learning. In this vein, extracting information from big data entails sorting, abstracting, valuing, and accumulating in order to establish a weighted or prioritized answer or set of answers. This is also a definition of information processing. The common definition of big data, simply means that vast amounts of data can be stored, retrieved and processed from the "cloud" (or network connected data stores).

If a human learns and can translate that knowledge and understanding to a machine, humans remain in control. Users of those machines can hold the humans that created the operational policies responsible for the outcomes. In many domains you may find experts that disagree. They disagree on the importance of factors that should be considered. They disagree on how influencing factors are integrated to make a decision or to control an action. In purely human systems it is often difficult to get a concise explanation of how they valued different factors in their decision-making. By precise, we mean a mathematically explicit explanation. On the other hand, if a machine makes a decision, or takes an action, it should be easy to get a mathematically explicit explanation, because machines work only on numbers: valued information.

Machine Learning without Human Learning (Singularity)

If machines learn and humans do not learn, then the machines are in control.

If machines learn, but cannot (or do not) explain how they think, humans are at the mercy of the machines.

At this time, machine learning systems are provided information by humans. The positive impact of this is that the machine is still under the control of humans. The negative impact is that the humans are imparting their own bias into the solutions, because they are selecting the information that the machine learning systems use to bias their decisions and actions.

Even if the human suppliers of information attempt to be unbiased, and knowledge is being extracted from “big data”, the results may be biased towards the most prolific experts. And with the ability of computers to generate information, it may be easy to influence the big data.

A purist’s definition of learning might be explained like this: a brick (machine with an empty memory) is thrown on a table. It knows nothing. Somehow this brick (system) absorbs all knowledge of the universe and is able to make all decisions and actions without human intervention. This concept may be as equally flawed as systems where humans provide the information on which to make decisions. This is because, right up to the point where our hypothetical system knows everything, it will be lacking some information that may allow it to make less than optimal decisions. One can look at how different humans learn different things and make different decisions. This may be another reason to isolate the knowledge capture process from the adaptive expert behavior delivery process. They have different timelines and risks.

Static Decision-making versus Adaptive Control

Speed is another reason to separate the knowledge gathering function from the decision-making component.

For example: It is probably not appropriate for your self-driving car to query a cloud based service for directions concerning how to avoid a pot hole in the road when, at the same time a child is running across the road to retrieve a rolling ball, and a weaving semi-trailer truck is approaching. In fact there should probably not be any extra time spent on retrieving information beyond the real-time information being provided by the active sensors. The available information should be adaptively valued and integrated by the local operational policy in order to control the real time maneuver required.

There are, of course, operational decision-making scenarios that can be termed static. These types of decisions can be made when there is time to query for relevant opinions from data in the cloud. So, there are still opportunities for system architects to review their objectives and choose the best technologies and solutions.

Thinking is Important

“Thinking is the process of using one’s mind to consider or reason about something.”[5]

[5] Google: thinking; https://www.google.com/#q=define+thinking

One might suggest that the purpose of "reasoning" is to establish the value or importance of information, and how influencing factors work together toward some objective.

When this process targets intelligence, or artificial intelligence, then, with thinking, we are developing guidance upon which to deliver decisions and/or actions. For humans, this might be stated as developing policies to guide the interpretation of information in order to make decisions or perform actions. We develop policies to guide humans in how to perform complex abstract tasks, because it would be impractical to develop explicit rules that described precisely how to address every problem that might be encountered. We trust (hope) that the humans have supporting ethics and background knowledge to fill in the blanks.

If our objective is to provide artificial intelligence in order to deliver intelligent expert behavior, then we need policies that incorporate a value system and define how influencing factors are integrated to make decisions or control actions.

If we expect machines to think completely on their own, then those machines must be able to create their own value system. They must identify (on their own) the influencing factors necessary to make decisions (or take actions), and identify (on their own) how the influencing factors work together to make the desired decisions and actions. Perhaps we are not there yet.

KEEL® Technology

Compsim's Knowledge Enhanced Electronic Logic (KEEL) Technology is a human driven expert system that makes it easy to capture, test, package, audit and explain complex adaptive policies that can be executed in software applications and machines. Humans provide the expertise by capturing their judgment and reasoning skills as operational policies for execution in a cognitive engine. The KEEL Dynamic Graphical Language (DGL) provides 1) a way to collect and test the policies before packaging for production, and 2) a way to audit and review decisions and actions made by the production systems.[6] The DGL makes it easy to visualize the value system and the information fusion model that drives the decisions and actions.

The DGL allows the complex (dynamic, non-linear, inter-related, multi-dimensional) problem sets to be developed with relative ease, without resorting to higher level mathematics or complex software coding. KEEL Tools (incorporating the KEEL dynamic graphical language) are used to auto-generate platform and architecture independent code in most common computer languages.

[6] NATO Guidance Document; Autonomous Systems – Issues for Defence Policy Makers, Chapter 9: Auditable Policies for Autonomous Systems (Decisional Forensics)"; http://www.compsim.com/Papers2014/Autonomous-Systems-Publication_Print.pdf

Summary:

If the real objective of artificial intelligence is to support ***machines delivering expert adaptive behavior,*** then KEEL Technology is available to accomplish that objective now.

By separating the creation of knowledge and expertise from the delivery of expertise in the form of expert adaptive behavior, it may be easier to focus on the true objective of AI.

As more decisions and actions are allocated to machines, many of those machines will become dependent on expert adaptive behaviors. Many of these decisions and actions will be complex. KEEL tools streamline the development of these complex, adaptive behaviors suitable for execution in machines.

KEEL Tools also support auditing and explaining of the KEEL-based decisions and actions (Explainable AI) through a process of “Language Animation”. Using Language Animation you can see the information fusion process and the valued information flow. You can see the systems *think*.

By separating the delivery of expert adaptive behavior from the development of knowledge or expertise, one can focus on delivering more with less. This technology can be implemented now. It will keep humans in control and (with explainable AI) problems can be fixed when they are identified. Humans will be responsible; not the machines.

Glossary:

Artificial Intelligence: machines delivering expert adaptive behavior

Behavior: the way in which a natural phenomenon or a machine works or functions

Creativity: the production of original concepts

Expertise: the use of judgment and reasoning in the delivery of services (beyond just following rules)

Intelligence: captured expertise

Judgment: the ability to incorporate a value system in making decisions (collective information processing)

Learning: the ability to acquire knowledge

Logic: sequential processing of information

Knowledge: map of valued information and relationships

Machine Learning (today): accumulating provided information

Planning: reviewing information and selecting most appropriate actions

Point Decisions: Decisions where there is time consider your options as in long range planning / resource development

Policies: general guidance on how to address problems without describing all of the details

Problem Solving: reviewing information making decisions

Profiling: pattern matching

Reasoning: integrating valued information

Rules: defining logic

Self-awareness: incorporating self value into the decision-making process

Understanding: capable of providing explainable decisions and actions