Knowledge
Enhanced
Electronic
Logic

Topic Area: General - Regarding Trustable AI

General Comment: Many of the discussion topics in the "Executive order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence issued on October 30, 2023" suggest bias toward the approach to AI based on Machine Learning (ML), Supervised Learning (SL), Deep Learning (DL), Large Language Models (LLM) and similar approaches where a machine is taught to learn from observed patterns in the hope that the machine can then integrate those taught patterns to make good decisions. Similarly, there seems to be a recognition of some problems that might exist relative to this approach. Among these are potential hallucinations (generating faulty information or results), embedded hidden bias, and unethical behavior. Any reference to the need for "Trustable AI" may be the term that encompasses concerns for these problem areas.

ML/SL/DL/LLM:

I would term these ML/SL/DL/LLM approaches to AI as "data-driven" approaches. By itself, this is not bad, but perhaps it is simply accepting the fact that the problems are now too complex for humans to address by themselves. This Machine Learning approach also somewhat mimics the way humans learn through observations and training. Machines can just perform this learning function faster than humans.

Learning from data points does, however, raise some concerns. The random, inconsistent, and inappropriate behavior of some humans may represent the same concerns with the ML/SL/DL/LLM systems. And remember that by giving responsibilities to machines there is the potential to mass-produce undesired behavior. Individual humans might be prosecuted or dismissed for their inappropriate behavior, but not machines.

Historically, humans build machines to do exactly what is desired. If they did not perform as desired, they could easily be fixed. If these were 20th-century intelligent machines, they were computer-controlled machines that followed explicit rules. In the past, humans remained responsible and were in control of the behavior of the machines. One could look inside and see exactly why the machine behaved the way it did. Nobody would consciously add random behavior to a factory machine.

One might suggest that today's description of Artificial Intelligence is an extension to 'past intelligent machine behavior' where the machines would take on the additional judgmental tasks that have required humans to perform in the past. It was the judgment and reasoning capabilities of humans that separated them from the intelligent machines of the past. With our constant demand to do more with less, there is a demand for Artificial Intelligence to handle the rapid increase in complexity and the amount of information that needs to be processed.

Algorithms:

Newtonian calculus is an approach used to define functional relationships between information items. So, one approach to pursuing the objectives of AI is to write the Algorithms for AI. This keeps humans in the loop, and they are responsible. And if appropriate reviews are in place for safety and quality, the ML/SL/DL/LLM issues may be eliminated. The problem with this approach may be one of economics. The domain experts may not be fluent in higher-level mathematics. There are many challenges translating concepts that are described in text, then converted into formulas, then converted into code that must be debugged and packaged before anyone can see if the results had anything to do with what was originally desired. Perhaps this time-consuming, costly process (which was dependent on certain skilled individuals) led to the ML/SL/DL/LLM approach to pursue the AI objectives. Yet the issues concerning ML/SL/DL/LLM still remain.

An Alternative Approach to AI using KEEL Technology:

There is another way to give machines the ability to apply judgment and reasoning skills to adaptively make decisions and to perform adaptive operational control. The complex, costly and time-consuming approach of manually developed algorithms can be made simple and easy without any dependence on higher level mathematics and skilled programming, there is an alternative to the ML/SL/DL/LLM approach without the ML/SL/DL/LLM baggage.

Knowledge Enhanced Electronic Logic (KEEL®) "Technology" makes it relatively easy to allow human domain experts to tell machines how to think, adapt, and behave. This is accomplished using the KEEL Dynamic Graphical Language (DGL) which has no prerequisite math or programming skills. At the same time, KEEL technology allows the machines to explain their behavior using a process called "Language Animation" at no additional cost. This makes it easy for humans to audit the behavior of the machines. KEEL Technology provides 100% Explainable and Auditable behavior wherever it is used.

What appears to be missing from the ML/SL/DL/LLM approach is an exposed value system. One might suggest that concerns regarding ethics and hidden bias are simply highlighting the need for an exposed value system. We might suggest that it is the values behind all factors in a decision or action that are important, not just those that might be associated with ethics or bias associated with specific objectives. To adequately evaluate the behavior of any decision or action requires one to identify the value assigned *to any factor, at any point in time*. This is provided with KEEL Technology

Suggested Requirements for Trustable AI:

1. **Explainable AI:** Albert Einstein said: “If you can’t explain it simply, you don’t understand it well enough.” Lord Kelvin said: “If you cannot measure it, you cannot improve it.” If it cannot be explained, it cannot be trusted to deliver the desired behavior. This means that any AI must be able to answer these questions:
   - What were **all** the options considered, and how was each valued and supported? (This lists all the decisions or actions the system has available.) List with specific numbered values.
     - This will show the system’s priorities.
   - List **all** the influencing factors (information items: pros and cons impacting the decision-making space), and how was each valued? (Show list with “numbered values”.)
     - This will expose the value system of the autonomous system.
   - How were **all** the influencing factors integrated to control the decision or action? (Show this explicitly, along with intermediate value integration points without any ‘fluff’.)
     - This will expose in more detail how **all** influencing factors and the value system combined to make decisions. The designer of the system will have given the system a “value system” and told it how to integrate the influencing factors to make decisions and control actions that will be exhibited by the system.
     - The reviewer should be able to see how the system adapted over time, leading up to the decision, or qualified / quantified action.

2. **Easy Auditing**: It must be easy to audit the behavior of AI-directed decisions and actions by a human with limited training. It cannot depend on experience with higher-level mathematics or programming.
   - This is mandatory because if it is too difficult, it will not be done!

3. **Support for Complex Problem Sets:** The primary reason for AI in the first place is the need to support Complex Problems. These types of problems may have one or more of these characteristics: “Non-linear characteristics” where influencing factors need to be interpreted in non-linear ways. Example: Temporal factors of time and distance where the closer a factor gets, the more important it becomes. “Dynamic Characteristics” where influencing factors may be constantly in motion. “Interrelated Characteristics” where influencing factors may impact different parts of the problem set in different ways. “Multi-dimensional Characteristics” where decisions and actions may be important in different ways at different times, and decisions and actions in the present must consider the future.

4. **Complete Coverage of the Decision Map:** To deliver Trustable AI, the AI must provide complete trustable coverage of the entire decision landscape. By complete, we mean that there cannot be any gaps in controlled coverage. To deliver Trustable AI this means that there cannot be interpolated points between taught patterns that are not 100% controlled! This means that to support complex problem sets (as defined in 3. above) coverage must be executed in a formula; no matter how that formula is developed/derived.
5. **Support for Safety Critical Mass-Produced Systems:** There are numerous opportunities to utilize AI that are not “Safety Critical”. These decisions could be less important than more critical decisions and actions, where failure to perform as desired may lead to critical errors. Examples of non-critical applications would be creating images or music that might satisfy one audience more than another, or creating “fluffy” language based on LLM to introduce a topic to humans. In these cases, there is no “correct” solution and one is just looking for “acceptable” solutions. Safety-critical decisions could be those that impact life and death, or potentially decisions that make investment decisions. Self-driving cars that make decisions where there are no ”good” solutions and people will be killed. The car is deciding who dies: the passengers in the car or the bystanders. This would be an example of a safety-critical, mass-produced solution. Military, fully-autonomous, Unmanned Combat Aerial Vehicles that make life-and-death decisions is another example.
6. **Emergency Stop:** The concept of an Emergency Stop is well understood in the factory automation market. There is a recognition that things will go wrong, either through human carelessness, or sensor failure, or failure of some component necessary in the performance of the manufacturing process. It is common to have a red button that will shut down a cell. In the manufacturing arena, this may simply remove power to the system. In the “complex problem” space the problems can come from many directions. There should be a requirement to exert absolute control of the system. In the ML/SL/DL/LLM space this may require a wrapper around the system. However, in a complex problem space where multiple problems are being addressed collectively, this may be more difficult to implement. Within KEEL Technology there is the methodology that “No” overrides “Yes”, or “Impossible” overrides “Must”. This concept is implemented throughout the KEEL cognitive process where all information is processed collectively. This removes any ambiguity in the operational policies. No external wrapper is required.

Future Requirements:

We suggest that we are entering the next phase of automation where fully autonomous systems will be competing without any human-in-the-loop control.

This will add one very important new requirement. That will mean that the time required to fix (or extend) systems in the shortest time will define survival. New information sources will be identified and will need to be integrated throughout the systems-of-systems. New tools/weapons will be developed that need to be considered. These new capabilities need to be integrated into systems faster than the competition to remain competitive.

Summary:

KEEL Technology makes it easy to deliver 100% Explainable and Auditable AI. Explainable AI should be a requirement for any system that requires trust.

And, beyond the "trust" requirements, there are other requirements that may be important. KEEL Technology enables:

- Independence from any software libraries
- Independence from any specific hardware
- Explainable AI for any application
- Explainable AI for any System Architecture (embedded devices, sub-assemblies, systems of systems, intelligent agents, web services)
- Explainable AI for real-time control
- Very small memory requirement for KEEL Cognitive Engines (enabling application in widgets, or sub-assemblies of larger systems)
- Expert "Point" Decisions (Example: decision to shoot)
- Expert "Adaptive Operational Control" (Example: drive a car / control a drone)

KEEL Technology is not a research project. As a "technology", it provides a different methodology to deliver Artificial Intelligence. KEEL "tools" facilitate the delivery of the technology. KEEL Technology was completely developed and funded by Compsim and is covered by granted patents, copyrights, and trade secrets. https://www.compsim.com