What Metrics Should We Use To Measure Commercial AI?

Cameron Hughes (Northeast Ohio ACM Chair; cameronhughes@acm.org)
Tracey Hughes (Northeast Ohio ACM Secretary; tracey.hughes@neoacmchapter.org)
DOI: 10.1145/3340470.3340479

In AI Matters Volume 4, Issue 2, and Issue 4, we raised the notion of the possibility of an AI Cosmology in part in response to the "AI Hype Cycle" that we are currently experiencing. We posited that our current machine learning and big data era represents but one peak among several previous peaks in AI research in which each peak had accompanying "Hype Cycles". We associated each peak with an epoch in a possible AI Cosmology. We briefly explored the logic machines, cybernetics, and expert system epochs. One of the objectives of identifying these epochs was to help establish that we have been here before. In particular we've been in the territory where some application of AI research finds substantial commercial success which is then closely followed by AI fever and hype. The public's expectations are heightened only to end in disillusionment when the applications fall short. Whereas it is sometimes somewhat of a challenge even for AI researchers, educators, and practitioners to know where the reality ends and hype begins, the layperson is often in an impossible position and at the mercy of pop culture, marketing and advertising campaigns. We suggested that an AI Cosmology might help us identify a single standard model for AI that could be the foundation for a common shared understanding of what AI is and what it is not. A tool to help the layperson understand where AI has been, where it's going, and where it can't go. Something that could provide a basic road map to help the general public navigate the pitfalls of AI Hype.

Here, we want to follow that suggestion with a few questions. Once we define and agree on what is meant by the moniker artificial intelligence and we are able to classify some application as actually having artificial intelligence, another set of questions immediately present themselves:

- How intelligent is any given AI application?
- How much intelligence does any given AI application have?
- How much intelligence does an application need to be classified as an AI application?
- How reliable is the process that produced the intelligence for any given AI application?
- How transparent is the intelligence in any given AI application?

The answers to these questions require some kind of qualitative and quantitative metrics. Namely, how much intelligence does any given AI application have and what is the quality of that intelligence. Further, how could we congeal the answers to these questions so that they can be used to capture (in label form) the 'AI Ingredients' of any technology aimed at the general public?

The amount of education is often used as one metric for intelligence. We refer to individuals as having a grade school, high school or college-level education. Could we employ a similar metric for AI applications? Would it be feasible to classify AI applications in terms of grade levels? For instance, an AI application with grade level 5 would be considered to have more intelligence than a 4th grade AI application. Any application that didn't meet 1st grade level would not be considered an AI application and applications that achieved better than 12th grade would be considered advanced AI applications. But how could we determine the grade level of any given AI application?

A consensus set of metrics that could be passed on to the general public has not yet prevailed. In this AI hype cycle, if an application uses any artifact from any AI technique a vendor is quick to advertise it as an AI application. It would be useful to have a metric that would indicate exactly how much AI is in the purported application. We have this kind of information for other products e.g. how much chocolate is actually in the chocolate bar or how much real fruit is actually in the fruit juice being sold to the consumer. Exactly how much AI does that drone have? Or how much AI is

actually in that new social media application? The 'how much' question requires a quantitative metric of some sort and the grade of intelligence involved requires the qualitative metric. What if applications that claimed to be AI capable were required to state the metrics on the label or in the advertising? For example, A vendor might state: "Our new social media application is 2% level 3 AI!" This kind of simple metric scheme would help to mitigate AI Hype cycles.

In addition to characterizing the quantity and quality of the embedded AI, requiring a reliability metric like MTBAIF (Mean Time Between AI Failure) is also desirable. Stating how much intelligence is in an application and what grade level of intelligence is in an application provides a good start. However, the reliability of the AI (i.e its limits, tolerances, certainty, etc.) and a transparency metric that indicates the ontology, inference predisposition/bias, and type and quality of knowledge would give the user some real indication of the utility of the application.

Knowledge Ingredients

What if we could provide 'Knowledge Ingredient' Labels for our AI-based hardware/software technologies like those shown in Figure 1.

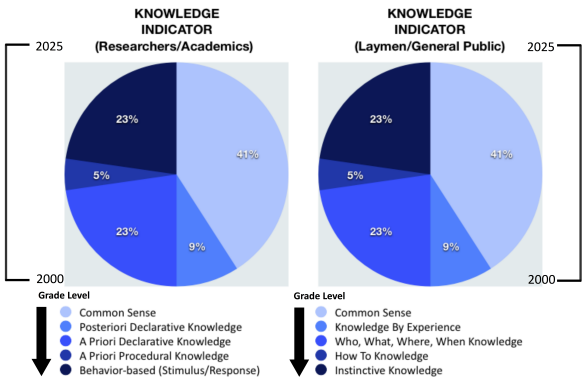

Figure 1: Pie charts for a Knowledge Indicator that reflects the Knowledge Ingredients of an AI system.

In Figure 1, the percentage of the types of knowledge the system is comprised of is represented in a pie chart. The color indicates the grade levels from lower to higher. In this case, the system has 41% of the knowledge is "Common Sense" with a relative low grade level as compared to the high grade level of "Instinctive Knowledge" at 23%. The expiration date indicates the time frame for the viability of the knowledge from 2000 - 2025. Outside of the indicated time frame, the knowledge would be consider obsolete. There are two indicators, one for researchers and practitioners and the other for the laymen. The difference in the indicators is the terminology used to describe the types of knowledge. Here is our lists of metrics:

1. Percentage of software/hardware dedicated to the implementation of AI techniques
2. Grade Level
3. Reliability (limits, tolerances, certainty, MTBAIF)
4. Transparency (predisposition/bias)

AI Metrics

The notion of metrics to measure AI performance has been under active investigation since the very beginnings of AI research and a standard widely used and accepted set of metrics remain elusive. The PerMIS (Performance Metrics for Intelligent System) workshops were started in 2000. The PerMIS workshops are dedicated to:

> *"...defining measures and methodologies of evaluating performance of intelligent systems started in 2000, the PerMIS series focuses on applications of performance measures to applied problems in commercial, industrial, homeland security, and military applications."*

The PerMIS workshops were originally cosponsored by SIGART (now SIGAI), NIST, IEEE, ACM, DARPA, and NSF. These workshops endeavored to identify performance intelligence metrics in many areas such as: ontologies, mobile robots, intelligence interfaces, agents, intelligent test beds, intelligent performance assessment, planning models, autonomous systems, learning approaches, and embedded intelligent components. ALFUS (Autonomous Levels For Unmanned Systems)

defines a framework for characterizing human interaction, the mission and environmental complexity of a ”system of systems”. The purpose of ALFUS was to determine the level of autonomy, a necessary but not sufficient component of an intelligent system. According to Ramsbotham [1]:

> *“Intelligence implies an ability to perceive and adapt to external environments in real-time, to acquire and store knowledge regarding problem solutions, and to incorporate that knowledge into system memory for future use.”*

describes the autonomous intelligent systems of systems based on the 4D/RCS Reference Model Architecture for Learning developed by Intelligent Systems Division of the NIST since 1980s. In order to attempt to develop some type of metric, the functions that comprised the intelligent behavior had to be decomposed into specific hardware and software characteristics. This was called SPACE (Sense, Perceive, Attend, Apprehend, Comprehend, Effect action):

- **Sense:**
  To generate a measurable signal (usually electrical) from external stimuli. A sensor will often employ techniques (for examples, bandpass filtering or thresholding) such that only part of the theoretical response of the transducer is perceived.
- **Perceive:**
  To capture the raw sensor data in a form (analog or digital) that allows further processing to extract information. In this narrow construct perception is characterized by a 1:1 correspondence between the sensor signal and the output data.
- **Attend:**
  To select data from what is perceived by the sensor. To a crude approximation, analogous to feature extraction.
- **Apprehend:**
  To characterize the information content of the extracted features. Analogous to pattern recognition.
- **Comprehend:**
  To understand the significance of the information apprehended in the context of existing knowledge–in the case of automata, typically other information stored in electronic memory.
- **Effect action:**
  To interact with the external environment or modify the internal state (e.g., the stored information comprising the ”knowledge base” of the system) based on what is comprehended.

The purpose or “collective mission performance” of a systems of systems was also categorized based on functional and architectural complexity. The mission and environmental complexity and the degree of human interaction determining the level of autonomy could be applied to these categories [1]:

1. **Leader-Follower**
   Intelligent behavior exhibited by single node, and replicated (sometimes with minor adaptation) by other nodes.
2. **Swarming (simple)**
   Loosely structured collection of interacting agents, capable of moving collectively.
3. **Swarming (complex)**
   Loosely structured collection of interacting agents, capable of individuated behavior to effect common goals.
4. **Homogenous intelligent systems**
   A relatively structured collection of identical (or at least similar) agents, wherein collective system performance is optimized by optimizing the performance of individual agents.
5. **Heterogeneous intelligent systems**
   A relatively structured heterogeneous collection of specialized agents, wherein the functions of intelligence distributed among the diverse agents to optimize performance of a defined task or set of tasks.
6. **Ad hoc intelligent adaptive systems**
   A relatively unstructured and undefined heterogeneous collection of agent, wherein the functions comprising intelligence are dynamically distributed across the system to adapt to changing tasks.

From less complex groups with low mission and environmental complexity and high human interaction (like Leader-Follower) to the most sophisticated high mission and environmental complexity and no human interaction (like Ad Hoc Intelligent Adaptive Systems),

these categories of systems are in a somewhat order. Figure 2 shows where the extremes of these categories could be graphed.

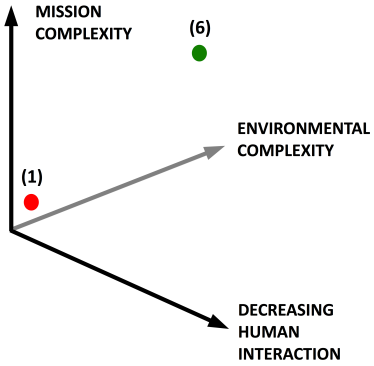

Figure 2: Lead Follower and Ad Hoc Intelligent Adaptive Systems are located on the 3D space of Complexity and Human Interaction.

Ultimately, Ramsobotham [1] concluded:

> *"Even given a comprehensive framework, as we begin to build more complex intelligent systems of systems, we will need to acquire knowledge and improve analytic tools and metrics. Among the more important will be: ... Better models and metrics for characterizing limits of information assurance based on these effects. This will be both a critical need and a major challenge."*

The most known metric used for researchers and commercial AI applications has been the Turing Test developed by Alan Turing in 1950. The purpose of the test was to evaluate a machine's ability to demonstrate intelligent behavior. That behavior was to be indistinguishable from a human being exhibiting the same behavior. A human judge was to evaluate a natural language conversation between a human and the potential "intelligent machine" shown in Figure 3.

Is this test actually a metric for "intelligence"? This has been debated for many years. If a NLG agent can produce responses that simulate human responses does that mean that the system is intelligent? Probably not, but that has not stopped the use of the Turing Test or AI software developers heralding that their

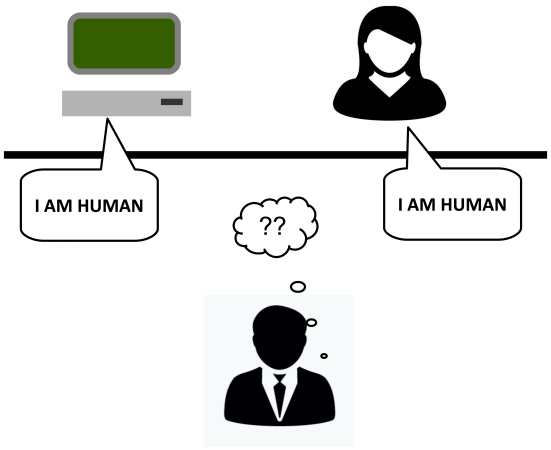

Figure 3: Human Evaluator was to determine which was not human.

systems has passed or almost passed the Turing Test. Now passing the test includes simulating the human voice like the Google Duplex AI Assistant.

In "Moving Beyond the Turing Test with the Allen AI Science Challenge" [2], the authors describe the Allen AI Science Challenge "Turing Olympics", a series of tests that explore many capabilities that are considered associated with intelligence. These capabilities include language understanding, reasoning, and commonsense knowledge needed to perform smart or intelligent activities. This would replace the Alan Turing pass/fail model. The idea of the Challenge was to have a four-month-long competition where researchers were to build an AI agent that could answer an eighth-grade multiple choice science questions. The AI would demonstrate its ability to utilize state-of-the-art natural language understanding and knowledge-based reasoning. Below summarizes the nature of the competition:

Number of total (4) choice questions: **5,083**

Training set Questions: **2,500**

Validation set confirming model performance: **8,132**

Legitimate questions: **800**

Final Test + validation set for final score: **21,298**

Final Legitimate Questions: **2,583**

Baseline Score random guessing: **25%**

The final results of the test, the top three competitors had scores with a spread of 1.05% with the highest score 59.31% which is considered a failing grade. Each of the winning model utilized standard information-retrieval-based methods which were not able to pass the eighth grade science exams. What is required is:

> *" ...to go beyond surface text to a deeper understanding of the meaning underlying each question, then use reasoning to find the appropriate answer."*

In this case, such a system would have a 1st grade level Knowledge Quality based on our quasi Knowledge Ingredient Indicator. Based on the article [2]:

> *"All three winners said it was clear that applying a deeper, semantic level of reasoning with scientific knowledge to the questions and answers would be the key to achieving scores of 80% and higher and demonstrating what might be considered true artificial intelligence."*

Here we've provided a very cursory look at a very limited set of possibilities for measuring commercial AI. In the next issue, will go a little further and dig a little deeper into the question of how do we communicate basic AI metrics and ingredients for commercial AI to the layperson.

References

[1] Ramsbotham, Alan.J. (2009). *Collective Intelligence: Toward Classifying Systems of Systems*. PerMIS'09.

[2] Schoenick, Carissa, Clark, Peter, T. et.al (2017). *Moving Beyond the Turing Test with the Allen AI Science Challenge*. CACM, VOL.60 (N0.9), pp. 60-64.

Cameron Hughes is a computer and robot programmer. He is a Software Epistemologist at Ctest Laboratories where he is currently working on A.I.M. (Alternative Intelligence for Machines) and A.I.R (Alternative Intelligence for Robots) technologies. Cameron is the lead AI Engineer for the Knowledge Group at Advanced Software Construction Inc. He is a member of the advisory board for the NREF (National Robotics Education Foundation) and the Oak Hill Robotics Makerspace. He is the project leader of the technical team for the NEOACM CSI/CLUE Robotics Challenge and regularly organizes and directs robot programming workshops for varying robot platforms. Cameron Hughes is the co-author of many books and blogs on software development and Artificial Intelligence.

Tracey Hughes is a software and epistemic visualization engineer at Ctest Laboratories. She is the lead designer for the MIND, TAMI, and NO-FAQS projects that utilize epistemic visualization. Tracey is also a member of the advisory board for the NREF (National Robotics Education Foundation) and the Oak Hill Robotics Makerspace. She is the lead researcher of the technical team for the NEOACM CSI/CLUE Robotics Challenge. Tracey Hughes is the co-author with Cameron Hughes of many books on software development and Artificial Intelligence.