Performance Metrics for Intelligent Systems

A. Meystel
Professor of Electrical and Computer Engineering
Drexel University, Philadelphia, PA 191004
alex@impact.drexel.edu

Extended Abstract

Is Testing of Intelligent Systems different from Testing of Non-intelligent Systems?

Testing of performance pertains to evaluation of the potential and actual capabilities of a system to satisfy the expectations of the designer and the users via exploration of its functioning. This includes determining how well the system performs its declared "job," how efficiently and effectively it does so, how robust it is, and so forth. The "job" and expected performance must therefore be defined at the outset. Efficiency is defined as how well the system does things right, effectiveness is defined as how well the system does the right thing, and robustness is defined as "the degree to which a system … can function correctly in the presence of invalid inputs or stressful environmental conditions." In general, the evaluation of intelligent systems (IS's) is broader than testing of non-intelligent systems (NIS). A system that has intelligence should in general be able to perform under a wider range of operating conditions than one that does not have intelligence. In addition, it should learn from its experiences and either improve its results within the same operating conditions or extend its range of acceptable conditions.

Function of Behavior Generation

Dealing With General and/or Incomplete Commands. An IS is given a job to do (task, mission, set of commands). *The job definition for IS is expected to be less specific than in an NIS*. A system with intelligence ought to have the capability to interpret incomplete commands, understand a higher level, more abstract commands and to supplement the given command with additional information that helps to generate more specific plans internally. The IS should understand the context within which the command is given. For example, instead of telling a mobile robot to go to a specific location in world coordinates "GO_TO(X, Y)," the command could be "Go to the window nearest to me." The robot should understand what a window is and know that it needs to find one which is the minimum distance away from me and move to that location

Ability to Synthesize the Alternatives of Decisions and to Choose the Best One. There was time, when the processes of decision making and planning were understood and reproduced as choosing from the preprogrammed lists and menus. This time has passed. Now, it became clear that most of the decisions should be synthesized on line. It becomes increasingly clear that most of the planning procedures require searching. It was discovered that the advantages of search algorithms can be achieved when the space is represented and search is organized in a multiresolutional fashion.

Function of World Modeling

Knowledge Representation. In most intelligent systems, an internal model of the world and/or a long-term knowledge store are utilized as a part of the overall knowledge representation system (KR). The long-term knowledge store (repository, or knowledge base) contains fairly invariant information, such as street maps or machining rules. An enabling aspect of the system's intelligence is the *a priori* knowledge it has and knows how to use.

The locally sensed information is obviously more current than that in the long-term store. Therefore, it must supercede what is in the knowledge base if there's a conflict. If a road has been closed, the system will plan around it and should, if appropriate, update the long-term maps. Obviously, these processes of updating our knowledge of the world belong to different levels of granularity, require different scale for interpretation and serve for supporting different resolutions of planning.

Commonsense knowledge. An intelligent system should be able to have generic models available that guide it as it interacts with the world. This is as opposed to non-intelligent systems, where the environment is constrained to fit within the system's expectations (limited knowledge about *what is possible*). Although all possible situations cannot be predicted, the system should be prepared to handle many of them by a sub-store of *commonsense knowledge*. For example, the system may have to recognize and model stairs and elevators if it needs to go between floors. Not all stairs have the same geometry or configuration. It must

know how elevators work, if that is appropriate to its job, namely, how to call an elevator, determine that one is available going in the right direction, selecting the floor, waiting until the right floor is reached and the door is open, etc. The intelligent system has to be able to map between the generic and the specific.

Processes of Knowledge Acquisition: Updating, Extrapolating, and Learning. The updating of all sub-stores is conducted as the new information arrives. This information is frequently incomplete as far as satisfying the documents and models used by IS. An intelligent system must also be able to fill in gaps in its knowledge. If a moving object appears behind a robotic vehicle, the vehicle notes that it has an unknown entity that must be identified. Is it an emergency vehicle that must be given the right of way or an aggressive driver? It has to extrapolate or interpolate based on what it knows and what it discovers.

Related to this is the concept of predicting what will happen in the future. A machine tool that has a model of tool wear should forecast when a particular cutter will need to be replaced. A mobile vehicle will have to estimate its own trajectory and that of others with which it could potentially collide. The multiresolutional planning processes use various horizons of anticipation (larger at lower resolution and smaller at higher resolution.

The ability to anticipate will be amplified by learning new phenomena and control rules from experience. An intelligent system should become better at performing its job as it learns from its experiences. Therefore, one aspect that should be part of the testing or evaluation is the evolution and improvement in the system's functioning.

Performance Evaluation in Numerical Domains

Requirements for Testing Intelligent Systems. Based on the discussion above, there is an initial set of requirements for testing intelligent systems that arise. The tests should therefore be designed to measure or identify at least the following abilities:

1. to interpret high level, abstract, and vague commands and convert them into a series of actionable plans
2. to autonomously make decisions as it is carrying out its plans
3. to re-plan while executing its plans and adapt to changes in the situation
4. to register sensed information with its location in the world and with a priori data
5. to fuse data from multiple sensors, including resolution of conflicts
6. to handle imperfect data from sensors, sensor failure or sensor inadequacy for certain circumstances
7. to direct its sensors and processing algorithms at finding and identifying specific items or items within a particular class
8. to focus resources where appropriate
9. to handle a wide variation in surroundings or objects with which it interacts
10. to deal with a dynamic environment
11. to map the environment so that it can perform its job
12. to update its models of the world, both for short-term and potentially long-term
13. to understand generic concepts about the world that are relevant to its functioning and ability to apply them to specific situations
14. to deal with and model symbolic and situational concepts as well as geometry and attributes
15. to work with incomplete and imperfect knowledge by extrapolating, interpolating, or other means
16. to be able to predict events in the future or estimate future status
17. the ability to evaluate its own performance and improve

Most of the items on the list allow for a numerical evaluation. However, non-numerical domains play a substantial role in evaluating intelligence and performance of IS.

Performance Evaluation in Non-numerical Domains

This theme focuses upon the aspects of intelligent system performance that are not directly quantifiable, but which should be subject to meaningful comparison. An example of an analogous aspect of human performance is the term "intelligent" itself. The notion of quantifying intelligence has always been controversial, even though people regularly use terms that ascribe some degree of intelligence. Terms ranging from smart, intelligent, or clever to dumb, stupid, or idiotic, with all sorts of degrees between, express people's judgments. But of course, these are often arbitrary judgments, without any basis for comparison or consistency of application. The notion of IQ, based on the widely used tests, was intended as a means of providing some consistency and quantification, but is still controversial.

So how might we do measurements for machines of the virtues that we associate with intelligence? First, we have to encapsulate the notion of what we mean by intelligence a little better. From the previous section one can see that the following properties are tacitly considered to pertain to intelligent systems:

- the ability to deal with general and abstract information
- the ability to deduce particular cases from the general ones
- the ability to deal with incomplete information and assume the lacking components
- the ability to construct autonomously the alternative of decisions
- the ability to compare these alternatives and choose the best one
- the ability to adjust the plans in updated situation
- the ability to reschedule and re-plan in updated situation
- the ability to choose the set of sensors
- the ability to recognize the unexpected as well as the previously unknown phenomena
- the ability to cluster, classify and categorize the acquired information
- the ability to update, extrapolate, generalize, and learn
- being equipped with storages of supportive knowledge, in particular, commonsense knowledge

Ontologies and Reasons for Comparing Them in Intelligent Systems

Whether an ontology is used within a computer program (or even the requirements statement of a planned computer program), a database (and its associated programs), a knowledge based system, or an autonomous artificially intelligent system, the ontology is indeed an informational core. As the architecture of the knowledge repository, the ontology (ontologies) are multigranular (multiresolutional, multiscale) in their essence because of multiresolutional character of the meaning of words. In integrating systems, the presence of a shared ontology is what will allows for interoperability. The term can be applied to the world-view of a human, too (in fact, is derived from a human study) though it may be easier to elicit it from the machine, as remarked above. (A fact related to the "knowledge acquisition bottleneck".) Thus it is an aspect of intelligent behavior that we may be able to compare from one system to another and correlate with the more general notion of intelligence in a system.

Returning to the best attempts to date to measure human intelligence, it is worth noting that a human's individual ontology might be explanatory for human intelligence, so it is not surprising that there are indirect measures of ontologies on IQ tests and achievement tests. To measure the breadth of the person's intelligence, is seems useful to ask if some people have "broader" ontologies than others. That is, do they cover more areas, or more subjects, or more aspects, or more details. Should we expect that these broader ontologies will manifest themselves in, say, a scholastic aptitude test (which in turn correlates with IQ)? Does the "broader" ontology testifies for the *breadth* of intelligence? Would that broader ontology influence the ability of the intelligent system (including humans) to make better decisions? For people, the answers seem to be "yes". It is tempting to imply that for machines, as well.

Humans use their ontologies (and actually, the whole system of knowledge representation) to label, categorize, characterize, and compare everything -- every object, every action. If a human learns the meaning of some new entity, it is because a label for this thing is put into the knowledge representation (KR) system, and eventually into a place in the ontology that relates it to the rest of the human's knowledge. If a human learns more about that entity, it is because more of its attributes, bounds, and relationships are specified in an Entity-Relational Network (ERN) of the knowledge representation (KR) where the ontology resides. The person does not have to bring all of its understanding of that same entity to conscious attention all the time, as it would be a distraction. So, the ontology is usually accessed only as much as needed to make the decision, or to communicate ideas and understand ideas communicated by others.

Decisions that lead to a high probability of success in dealing with the external world can only be made in the light of an individual's KR-based understanding of the facts surrounding the decision. If that individual does not have alternative actions characterized by information in an ontology, that individual cannot compare thee alternatives, and therefore cannot consider them in rational decision processes. If an organism's ontology does not reflect reality, the organism will make irrational and perhaps unsuccessful decisions. Complex decisions involve problem solving, and we must address methods of solving problems.

Measuring Non-Numerical Aspects of Intelligent Systems Related to Ontologies.

Can we exploit the idea of the human ontology above as a "core" of intelligence to characterize and compare intelligent behavior is machines based on a machine's ontology, built-in or acquired? Like a human,

a machine may have sensors connected to subsystems of sensory processing. The machine may be able to take certain actions that provide grounding for the ontology. If it can learn, perhaps it can extend its ontology. How can we characterize that ontology in a way that will allow us to characterize the machine's capabilities? How can we characterize its ability to change the ontology? If it has an ability to communicate to other machines or people, how does this ability add to its capabilities (and to its ontology)?

Evaluation: Mathematical and Computational Premises

Consider a general situation: there is a set of goals (G_1,..., G_n) and a set of IS (or intelligent agents) to achieve these goals. Different intelligent systems, or agents might have different goals, or they might put different weights on the various goals. Further, they might be better or poorer at pursuing those goals in differing contexts. That is, they might have different components of intelligence (I_1, I_2,..I_s) and these would be more or less important in the different contexts (C_1,..., C_q) that should also be known.

This dependence on the context determines that agents might be good at one set of matters, but bad in others. The agent might be good at trying and learning about recognizing new objects in the surrounding world, but poor at doing anything risky. It is typical for humans to have a portfolio of "intelligences" as well as "goals." It would give some value to all the different goals, and would have some value to each dimension of intelligence. One agent might be characterized as an explorer, while another is very good in performing repetitive routines. Which agent should be evaluated as a preferable one? Obviously, this would depend on the goal and the context. An unequivocal answer might be impossible at a single level of resolution because the true result depends on the distribution of the types of agents and the contexts that the groups of agents find themselves in. Thus, the "intelligences" as well as "goals" might require representing them as a multiresolutional system.

Evaluation of intelligence requires our ability to judge the degree of *success* in a multiresolutional system of multiple intelligences working under multiple goals. This means that if success is defined as producing a summary of the situation (a generalized representation of it), the latter can be computed in a very non-intelligent manner especially if one is dealing with a relatively simple situation. Indeed, in primitive cases, the user might be satisfied by composing a summary defined as "list the objects and relationships among them" i.e. a subset of an entity-relational network (ERN). On the other hand, the summary can be produced intelligently by generalizing the list of objects and relationships to the required degree of quantitative compression with the required level of the context related *coherence*. Thus, *success* characterizes the level of *intelligence* if the notion of *success* is clearly defined.

The need in determining levels or gradations of intelligence is obvious: we must understand why the probability of success increases because somebody is supposed to provide for this increase, and somebody is supposed to pay for it. Spatio-temporal horizons in knowledge organization as well as behavior generation are supposed to be linked with spatio-temporal scopes admitted for running algorithms of generalization (e.g. clustering). Indeed, we do not cluster the whole world but only the subset of it which falls within our scope. This joint dependence of clustering on both spatial relations and the expectation of their temporal existence can lead to non-trivial results.

Generalization (the ability to come up with a "gestalt" concept) is conducted by the virtue of recognizing an object within the chaos of available spatio-temporal information, or a more general object within the multiplicity of less general ones. The system has to recognize such a representative object, event, or action if they are *entities*. If the scope of attention is too small, the system might not be able to recognize the entity that has boundaries beyond the scope of attention. However, if the scope is excessively large, then the system will perform a substantial and unnecessary job (of searching and tentatively grouping units of information with weak links to the units of importance).

The Tools of Mathematics and The Tools of Computational Intelligence. Each of the tools discussed in this paper allows for a number of comprehensive embodiments by using standard or advanced software and hardware modules. Thus a possibility of constructing a language of architectural modules can be considered for future efforts in this direction.

Proper testing procedures should be associated with the model of intelligence presumed in the particular case of intelligence evaluation. It seems to be meaningful to compare systems of intelligence that are equipped with similar tools. In this section we introduce the list of the tools that are known from the common industrial and research practice of running the systems with elements of autonomy and intelligence. It is also expected that these tools can be used as components of the intelligent systems architectures. Thus,

they might help in developing and applying types of architectures that will be used for comparing intelligence of systems.

Learning. We have separated this into an independent sub-section because of the synthetic nature of the matter. Learning is the underlying essence of all phenomena linked with functioning of an intelligent system. It uses all mathematical and computational tools outlined for all other subsystems. In the machine learning community, the attention is paid to three metrics: the ability to generalize, the performance level in the specific task being learned, and the speed of learning. From the intelligence point of view, the ability to generalize is the most important since the other two capabilities dwell on the ability to generalize. Systems can do rote learning, but without generalization, it is impossible, or at least very difficult to apply what has been learned to future situations. Of course, if two systems were equivalent in their ability to generalize, with the same resulting level of performance, then the one which could do this faster would be better.

Further Perspectives of Performance Evaluation[1]

Technology Readiness as a Performance Measure.

Technology Readiness Levels (TRLs) were initially proposed by NASA in 1995. The sporadic use within the US Science and Technology Community followed. They were adopted by the US DoD in June 2001 where they are now mandated for all major Acquisition programs.

The stages of technology readiness might seem to be a check-list for the stages of system's research, development, design and manufacturing. In fact, this is much more than just a reminder of not skipping the required stages. It is possible to demonstrate, having all these stages completed properly, testifies for the overall quality of the system. Indeed, each consecutive Technology Readiness Level (TRL) demonstrates at a higher resolution more narrow scenario of functioning.

Uncertainty and Complexity in the Environment

One of the problems encountered by any intelligent system is dealing with uncertainty and complexity in the environment. This may include the geometric and dynamic uncertainty and complexity. It might involve the number and type of moving objects. It might involve the nature of other agents within the environment. How intelligent are the other agents? Are they friendly or hostile? What are their physical capabilities? What are their intentions? How can these parameters be measured and quantified?

There are a number of ways to approach this problem. For example, control theory addresses questions such as, "How much information about a system's input-output behavior is needed to control it to a specified accuracy? How much identification is required if only rough bounds on time and frequency responses are available a priori." G. Zames links the cost of adequate control for imprecisely modeled systems with the value of the complexity of information processing required to achieve a prespecified accuracy. Zames suggests that Kolmogorov's ε-entropy is a better measure of complexity than anything else.

Kolmogorov was inclined to view entropy as a measure of complexity rather than information, (as Shannon did). Uspensky reflects on ε-entropy as follows: "complexity of things (as opposed to the complexity of processes, e.g. of computational processes) took the name descriptional complexity, or Kolmogorov complexity. Kolmogorov was taking in account objects and encodings of objects. The complexity of an object is the minimal size of its encoding.

Kolmogorov complexity has proven to be useful for evaluatiin of encodings (approximations) of functions specified to a particular precision, ε. The approximation of functions using lower dimensional subspaces has been explored extensively, and developments reported in numerous sources demonstrate their applicability to the measuring of performance of the intelligent systems.

Kolmogorov presents his version of entropy as being "of interest in the non-probabilistic theory of information in the study of the necessary size of memory and the number of operations in computational algorithms." Using e-entropy for complexity evaluation was demonstrated for a multiresolutional intelligent system. We can anticipate that by using computational complexity as one of the performance measures we can improve the existing system of performance evaluators (metrics).

---

[1] This section was written with participation of J. Albus (NIST) and E. Messona (NIST)

Psychophysical and Biometric Approaches. Many psychophysical experiments are designed to measure the ability of biological intelligence to analyze information in the presence of uncertainty and complexity.

Two categories of biometric techniques should be taken in consideration: physiological based and behavior based. Physiological based techniques measure physiological characteristics such as patterns in fingerprints, the iris, facial characteristics, geometry of the hand, vein patterns, the shape of the ear, body odor, DNA analysis, and sweat pore analysis. The behavioral based techniques measure the parameters such as: handwritten signature analysis, keystroke analysis, and speech analysis.

There are two basic concerns in these technologies: the error tolerance and the storage of the templates. The error tolerance of these systems is critical to their performance. Both errors (False Rejection and False Acceptance) should be low, and they should both be determined together with the manufacturers of sub-systems (components).

The recorded biometric measurements of a user (templates) can be stored in various places depending on the application and the security requirements of this application. The templates can be stored in the biometric device, in a central knowledge base or in portable carriers..

Reliability and acceptance of a security system depends on how the system is protected against threats and its effectiveness to identify system's abuses. There are various sources of threats that the biometric technologies face.

Linguistic approaches. One of the characteristics of intelligent systems is that they communicate (exchange messages) with each other. Therefore, the question of how to measure the performance of communication between intelligent systems arises. What is communicated? How is the information encoded? What is the bandwidth required? How useful is the information to the sender and the receiver? How secure is the channel (can communications be prevented from being intercepted by unwanted listeners?). Focus upon Natural Language proficiency arises in a natural way because of the communication issue. On the other hand, the decision making processes often benefit from being equipped by an ability to deal with Natural Language structures.