Standardizing Measurements of Autonomy in the Artificially Intelligent

Amy R. Hudson
National Institute of Standards and Technology
University of Maryland College Park
Gaithersburg, MD
ahudson5@umd.edu

Larry H. Reeker
National Institute of Standards and Technology
Gaithersburg, MD
larry.reeker@nist.gov

Abstract— The amount of control that an intelligent system has over their actions, whether they are able to act independently from their creator, plays a major factor in describing systems and in distinguishing them from each other. Different levels of autonomy reflect the different abilities of the machines as well as where and how they can play a part in our daily lives. We may begin to comprehend these abilities and possible applications into human society once we can classify the levels of autonomy. The goal of this project is to set a framework for establishing a standard of autonomy in the scientific community by examining past and current methods of measurement as well as exploring different levels of autonomy's ethics and implications. Once this framework is made available to the scientific community, more tests and experiments will be conducted to refine and further ingrain the framework so that classifications of artificial intelligent agents are available universally. If we are to continue improving upon our machines, developing them to be more adept at communicating and accomplishing tasks, then a set of standards must be established for the safety and convenience of mankind.

Keywords: *autonomy, artificial intelligence, standards*

Artificial Intelligence (AI) is a relatively new area of research and study, having been around for only fifty years. Its waters are uncharted, and as time progresses, there are more and more possibilities of research brought to the foreground, a multitude of things we may accomplish in collaboration with humans and the artificially intelligent. DARPA (Defense Advanced Research Projects Agency) is extremely interested in employing robots to take over the more dangerous and dirty parts of being a soldier, such as searching for enemy combatants in a crowd or unknown territory, to fetching injured soldiers from the battlefield. The threat of quivering in surgery has become obsolete with the use of robots to help out in making precise incisions into patients, just as countries with aging populations are considering the benefit of having robots available to care for the elderly. Through communicating when pills should be taken, providing or aiding in transport, the burgeoning AI generation will care for our elderly, taking our dispositions on retirement and our progressing age and making them more pleasant and convenient to both the aging and the people who have to care for them. This lightening the load, so to speak, captures the drive behind further investigating AI and using it as a source of helping society to cope with unpleasant tasks, as well as providing a new curiosity with which its citizens are free to explore and philosophize about. Now that the importance of AI research and testing has been (at least tentatively) established, we can proceed to emphasize the importance of having standards of measurement for artificially intelligent machines, both expediting and simplifying the research process, going about in an organized way so that the AI world will be on the same page and better prepared to provide insights that can easily be jumped off from to reach a new idea and new way in which our community may benefit.

The authors feel that the most pressing part of a machines makeup to standardize is their autonomy level. Autonomy (or the closeness of it) is THE major factor in determining what task a robot is able to complete; knowing how much human supervision and maintenance provisions a company will have to provide for a machine can help the company in either budgeting for these provisions, or perhaps investing in developing a program for the machine that allows it to conduct its task in its environment with much less human intervention, and therefore less costs. If autonomy is standardized then companies across the world would be able to accurately gauge expenses, and computer scientists interested in AI and robotics could test their latest exploits in making a machine more independent by using the standards both their competition is using, and the same standards they themselves have used in the past. There are already a wide variety of machines relying different amounts on their human creators, from vacuums that sense surfaces and clean your house without you pushing them or directing them, to cars that are able to navigate across deserts. With standard autonomy levels we will have a convenient way of measuring improvement in the field of AI.

There are five areas in particular where a human may intervene on the behalf of a machine. If these five areas summarize autonomy levels, then measurements of these areas can then be translated into a general measurement of autonomy, even as standardizing these measurements would standardize measurements of autonomy. To accurately standardize these measurements, research must be done on the spectrum of abilities our machines have now, matching today's autonomy levels with levels we can deduce after learning more about AI's possibilities in future research. This paper provides a framework which may prove useful to the scientific community when they are compiling these autonomy levels [see Epilogue below].

The first three areas where a machine may require aid from a human will be relatively simple to measure, especially if an ontology of autonomy is created amongst robots, so that the same scale and checking box can be used for every robot. Whether a robot can replicate, or can change programming or physical aspects of itself to better suit what it interprets its task to be, reflects its creating abilities. A one would be given to the system that can only be created with human help, a two to the system that can only create software and program or the system can only recreate physically, and a three to the system that is able to replicate itself completely; these numbers are just suggestions. A robot could be dependent on humans or other robots for movement to complete its task, being worn or carried perhaps. Is the system able to move without human interaction? The less independence it has, the lower score. Same thing for maintenance; is the system able to fix any malfunctions, physical and/or technical? Robots can get damaged in the course of working and a company that invests in a robot that is able to fix glitches in its system will save money. Also, if new programming updates are constantly being added by the robot automatically then the technology will always be up-to-date, and the robot will be able to accomplish more, faster! The machine must continue to adapt to its environment in order to be truly autonomous.

Communication is often an area where a human could intervene, maybe always present to translate what the robot senses into terms that they are able to, if not understand, then at least use to complete their tasks. Or the human (or another robot) would have to explain to other humans or robots what this machine is trying to show. Is the system able to effectively communicate with other machines and other humans? For measuring communication autonomy, the robots and people that the machine interacts with may be polled, with 50% of robots understanding the machine all the time, and only 10% of humans understanding it (without a "translator"). If the majority of the interactants do not understand the machine at anytime, the machine has the lowest communication autonomy, whereas if the majority of interactants understand all the time, then the machine gets the highest score. Surveys and questionnaires may aid in providing the researchers with the opinions of the surrounding public that has interacted with the machine, and would need to be modified as our communication trends are constantly being altered. Similar to communication, society's views on learning and testing for understanding vary significantly across both time frames as well as different cultures. Therefore, standardizing these components of autonomy should prove challenging.

The learning component of autonomy is one of the harder facets to measure quantitatively. We have created a general scale to organize the learning levels on, shown below:

1. Human solves problem with computational and data management

2. Computer interface adapts to human preference interactivity

3. Human specifies list of items and attributes and system clusters them

4. Supervised Learning: Human gives exemplars and computer learns

5. Reinforcement Learning: human interacts in learning by providing evaluation functions for system outcomes at various places

6. Means-ends analysis

7. Computer learns and transfers knowledge

8. Computer learns and can deal with new environment

This scale still leaves learning with an open-ended definition, one which different readers will interpret to mean various things. In the study of artificial intelligence, we are inflicting machines with our intelligence. It follows that these machines should be able to be tested on their learning skills with the same methods we employ on ourselves to reflect exactly what the person taking the test is capable of learning. Humans often do not completely understand each other (syntax and diction getting in the way), making analysis of tests shaky as answers need to be interpreted accurately. By participating in human discourse and grasping a greater understanding of it, we may develop AI systems that can communicate with humans and learn from them even better than humans can, as a human could adapt to the machine and then the machine might respond by adapting to the human.

Many studies have been conducted using the Structure of Observed Learning Outcomes (SOLO) taxonomy to determine how advanced a person's ability of learning and analysis is. The essay structure section of the analysis in SOLO is not applicable to our machine situation, but just as essays need to be well-ordered and structured, with lots of examples and specifics, machines the most independent from their creators will have an organized programming system and be able to create and interpret programs in that same organized fashion. English teachers are trained in reading a paper and knowing exactly which grade it deserves, due to how they have to sit and grade a multitude of papers at once. If English has found a way to change a qualitative assessment into a quantitative one, programmers should have training of assessing the structure and content of programs included in their education repertoire to accomplish a similar feat.

1. Misses the point
2. Single point
3. Multiple unrelated points
4. Intermediate
5. Logically related answer
6. Unanticipated extension

There are also many different types of learning, so that reinforcement learning, case-based learning, and others, could be assessed and compiled to all aid in reflecting the general learning capabilities of the machine. The scientific community will have to determine through experimentation how many of the different learning techniques should factor into the general learning assessment, and which of the many techniques even apply to learning in machines. Using only one or two of the known methods of testing learning may still be enough to obtain the score and level that is needed to interpret autonomy. If the accuracy of these tests are also recorded, then using just these tests are good enough for our purposes…is there really going to be a major difference between having an autonomy score of 3 and 3.5? and what will it be? Perhaps the standards of autonomy should become satisficed, since a more general term of autonomy will be more understandable to the public and maybe mean more for future calculations.

Once all of these five aspects of autonomy have been measured, they then can be added together to obtain another number. Sheridan's Model can then be broken up into appropriate number ranges for a 1 level of autonomy etc., so that the number you obtain from the five aspects of autonomy falls in a range for a specific level of autonomy.

Sheridan's Model: Levels of Autonomy in Decision Making [5]

100 % Human Control

1- Human considers decision alternatives, makes and implements a decision.

2- Computer suggests set of decision alternatives; human may ignore them in making and implementing decision.

3- Computer offers restricted set of decision alternatives; human decides on one of these and implements it.

4- Computer offers restricted set of decision alternatives and suggests one; human may accept or reject, but decides on one and implements it.

5- Computer offers restricted set of decision alternatives and suggests which one it, the computer, will implement if human approves.

6- Computer makes decision; necessarily informs human in time to stop its implementation.

7- Computer makes (implements) decision; necessarily tells human after the fact what it did.

8- Computer makes and implements decision; tells human after the fact what it did and only if human asks.

9- Computer makes and implements decision; tells human after the fact what it did only if it, the computer, thinks human should be told.

10- Computer makes and implements decision if it thinks it should; tells human after the fact if it thinks it should.

100 % Computer Control

The ranges of numbers to assign for the levels of autonomy also fall into the category of research and testing that needs to be conducted to obtain a solid ground for standardization, which may be updated as the years progress and more machines are created and not able to be labeled as a specific level. Updates to these standardizations will make for a much more robust way of measuring autonomy, being able to fix qualitative data to quantitative data so that much less bias is involved in the measurements.

One way in which we propose to instill these standards of autonomy in the future is to modify intelligent systems' ontologies to include an autonomy scale, created based off of the aforementioned qualities of learning, communication, movement, maintenance and creation. If the machine is able to rate itself on the scale, it will be able to communicate to people how much of a task it can be entrusted to perform without human input; how reliable it is. With this knowledge, people will be able to specify the best purpose the machine would serve, enabling the progression of forays into the AI realm of thought to be both organized and efficient.

EPILOGUE: SIX YEARS LATER: AUTONOMY & DEEP MEASUREMENTS FOR SCIENTIFIC CONSTRUCTS

We have made some suggestions on measuring autonomy in this paper, since it was a theme in this year's PerMIS, as well as being widely used today. “Theoretical Constructs and Measurement of Performance and Intelligence in Intelligent Systems” [Reeker, 2000] had a different role of trying to use measures and discover how they might help in developing scientific theories (e.g. Archimedes finding a way to measure amounts of gold and silver in a crown resulting in the study of hydraulics). In the 2000 paper, we also briefly talked about robustness and autonomy in extraction of information from text.

Our discussions to this point in this paper are more about measurements of autonomy in the engineering fields rather than a precise measurement leading to a part of a scientific theory. In the 2000 paper, engineering (by itself) is classified as a science of the artificial (certainly true with robotics). But engineering and pure science are linked in many areas, as one can see computational and scientific theories in the work of Herbert Simon and Allen Newell*. Testing for Autonomy by using levels of Autonomy or readiness levels and finding out how the system performs for each level, in what can be called a suite of performance metrics, specifies the machines abilities. Of course it doesn't tell whether the system is truly intelligent, and it probably would not satisfy Lord Kelvin, for instance. One can have multiple suites and multiple requirements for readiness, the different uses given in a suite on particular tasks.

The idea of the 2000 paper was to show that a young science like artificial intelligence should be looking for measures that are not only useful in developing technologies but can help put together a new science. This new science will have theories that then can be predictive through a calculus - or, as Simon and Newell had advocated - a computing program of the sort that is called today “computational science”. Such theories tend to be able to predict what is going to happen in various situations that stand up within empirical experiments, and it was called in that paper “deep measurements” (instead of “surface measurements”). In fact, the importance of being able to develop computational systems as part of a theory is of interest to many of us, and computational science continues its ascent.

We most often see computational science used in [Bekey 2005] and [Mataric, 2007] and not only in robotics, but also in The University of Massachusetts' Autonomous, Learning Laboratory's all types of learning. It is often used in discussing systems of agents, which can encapsulate information -- for Simon, “the allegory of the watchmakers” in [Simon, 1969].

So what is in this concept of autonomy? In the first part of this paper, we talked about human interaction with autonomous systems, for robots and other systems. In the spirit of the 2000 paper, we would like to look at aspects of autonomy that might be deeper: Robustness (already mentioned in the 2000 paper), stability, adaptability, capability, and scalability. Stability is associated with control theory. Its usefulness was around in physical systems long before AI, but some new ideas for the information age can be found at e.g. [Reeker, L.H. and Jones, A.T., 2001].

Adaptability is one of the most interesting of the “ility”s and is changing the world because information can be sent so quickly, whether it is in fly-by-wire systems in airplanes to the world-wide web and cell phones. In AI systems, we see lots of these adaptability possibilities. In the 2000 paper, these ideas have been discussed through learning and transfer of learning. The ideas of machine learning are getting better, and testing is important to them. For example, the ensemble methods have shown how to take learning systems and make interchanges between variance, noise (e.g. outliers), and bias. They can be tested using ROC charts (which are hardly different) that tell likely false positives and false negatives. New ideas have emerged for finding information in great amounts of text, and it has been shown that both ROCs and the standard precision and recall measurements used together are better than only one of either.

Ontologies, which were only mentioned in the 2000 paper, are clearly needed for communication between intelligent systems, whether human intelligence is studied by cognitive systems or artificial intelligence systems. Ontologies need more study and they are getting it through different efforts, such as the Semantic Web. They need tests and metrics, too; PerMIS is doing an important job in developing computer and information science and changing the world.

*Many people do not realize the amount of philosophy that both Allan Newell and Herbert Simon used in their work. They are thought of as renaissance men, mostly as in computer science, mathematics and statistics, cognitive science, and more, but they were also both philosophers, important in helping to develop the new science of artificial intelligence.

To conclude, there is still a lot of need for metrics of all types, both at the surface and in depth.

REFERENCES

[1] Bekey, G.A., [2005], Autonomous Robotics, MIT Press

[2] Mataric, M. J. [2007], The Robotics Primer, MIT Press

[3] Reeker, L.H. [2000], Theoretic Constructs and Measurement of Performance and Intelligence in Intelligent Systems, Proceedings of the 2000 PerMIS Workshop. Available from NIST/MEL

[4] Reeker, L.H. and A. Jones, [2001], Measuring the Impact of Information on Complex Systems, Measuring the Performance and Intelligence of Systems: Proceedings of the 2001 PerMIS Workshop. <http://www.isd.mel.nist.gov/research_areas/research_engineering/Performance_Metrics/past_wkshp.html>

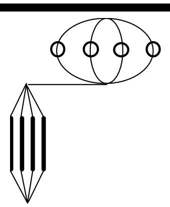
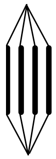
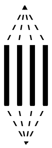
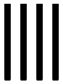
[5] R. Parasuraman, T.B. Sheridan, and C.D. Wickens, "A Model for Types and Levels of Human Interaction with Automation Transactions on Systems, Man, and Cybernetics- Part A, vol.30, pp. 286-297, 2000

[6] Simon, H.A., [1969], The Sciences of the Artificial, MIT Press

[7] The Chinese University of Hong Kong, <www.cuhk.edu.hk/clear/download/PDC/n23_SOLO_assessmt_grid.doc>

Appendix A:

The SOLO taxonomy as a guide to setting and marking assessment

SOLO category	Representation	Type of outcome	Solution to problem	Structure of essay
Unanticipated extension		Create Synthesise Hypothesise Validate Predict Debate Theorise	Solution to problem which goes beyond anticipated answer. Project or practical report dealing with real world ill-defined topic.	Well structured essay with clear introduction and conclusion. Issues clearly identified; clear framework for organizing discussion; appropriate material selected. Evidence of wide reading from many sources. Clear evidence of sophisticated analysis or innovative thinking.
Logically related answer		Apply Outline Distinguish Analyse Classify Contrast Summarise Categorise	Elegant solution to complex problem requiring identification of variables to be evaluated or hypotheses to be tested. Well structured project or practical report on open task.	Essay well structured with a clear introduction and conclusion. Framework exists which is well developed. Appropriate material. Content has logical flow, with ideas clearly expressed. Clearly identifiable structure to the argument with discussion of differing views.
Intermediate			Solution to multiple part problem with most parts correctly solved but some errors. Reasonably well structured project or practical report on open task.	Essay fairly well structured. Some issues identified. Attempt at a limited framework. Most of the material selected is appropriate. Introduction and conclusion exists. Logical presentation attempted and successful in a limited way. Some structure to the argument but only limited number of differing views and no new ideas.
Multiple unrelated points		Explain Define List Solve Describe Interpret	Correct solution to multiple part problem requiring substitution of data from one part to the next. Poorly structured project report or practical report on open task.	Essay poorly structured. A range of material has been selected and most of the material selected is appropriate. Weak introduction and conclusion. Little attempt to provide a clear logical structure. Focus on a large number of facts with little attempt at conceptual explanations. Very little linking of material between sections in the essay or report.
Single point		State Recognise Recall Quote Note Name	Correct answer to simple algorithmic problem requiring substitution of data into formula. Correct solution of one part of more complex problem.	Poor essay structure. One issue identified and this becomes the sole focus; no framework for organizing discussion. Dogmatic presentation of a single solution to the set task. This idea may be restated in different ways. Little support from the literature.
Misses the point			Completely incorrect solution.	Inappropriate or few issues identified. No framework for discussion and little relevant material selected. Poor structure to the essay. Irrelevant detail and some misinterpretation of the question. Little logical relationship to the topic and poor use of examples.