Proceedings of the 8th IEEE Mediterranean Conference
on Control and Automation (MED 2000)
Rio, Patras, GREECE 17-19 July, 2000

Panel Discussion

Measuring Performance of Autonomous Intelligent Systems: An Open Fundamental Problem

Co-Chairs/Organizers: **Panos Antsaklis, Alex Meystel**

Chair/Moderator: **Spyros Tzafestas**

Panelists (in alphabetical order):
Panos Ansaklis
Tamer Basar
Alex Meystel
George Saridis
Spyros Tzafestas
George Vachtsevanos

Discussion participants:
Vladimir Pilishkin
Elpida Tzafestas
Holger Voos

Thousands of person-years have been devoted to the research and development of various aspects of intelligent control systems with elements of autonomy. Much progress has been made. However, there exists very little ability to quantitatively evaluate that progress. The study of intelligent control systems is not a single field of study, but rather many fields, ranging from control theory and neural nets to artificial intelligence and cognitive sciences. How can we assess the current state of the science and technology under conditions that much diversified? Some systems are beginning to be deployed commercially. How can a buyer evaluate the advantages and disadvantages of different systems and decide which will perform best for any specific application ***if the intelligence and autonomy are required***? If one wishes to construct a system from existing components, how does one select those modules that are most appropriate for the target application?

Having the ability to measure the capabilities of intelligent control systems and their components is more than an exercise in intellectual curiosity or cognitive philosophy. To paraphrase William Lord Kelvin, when you can measure something and put some numbers to it, then you know something about it; and if you cannot, your understanding of it is of a "meager and unsatisfactory kind." Without metrics, it is a very difficult to quantify progress, evaluate results, reward success, or punish failure. It is therefore in a spirit of both scientific inquiry and pragmatic investigation that we are going to discuss ***metrics for evaluating the performance of intelligent systems***.

Many researchers have suggested measures of performance for intelligent machine systems. Perhaps the most famous is the Turing test [1]. Turing proposed that if a human using a keyboard cannot tell whether it is a human or a machine on the other end of an electronic conversation, then the machine is intelligent. There are two important features to this test. The first is that human performance is the metric against which machine performance is compared. The second is that the test is limited to language text interpretation. Turing's test does not measure of the ability of a machine to perceive images, to perform tasks, to plan and execute action, or even to understand spoken language. Newell expanded the list of abilities that a system must have to qualify as intelligent2. He proposed the following list [2]:

- recognize and make sense of a scene
- understand a sentence
- construct a correct response from the perceived situation

- form a sentence that is comprehensible and has a meaning of the selected response

- represent a situation internally
- be able to do tasks that require discovering relevant knowledge

This definition requires sensing, modeling, and output. However, it does not specify what "tasks" the machine must be capable of performing. This suggests that intelligence may be domain specific. Thus, situations and actions and the ability to perform tasks may vary with changes in the environment.
Albus has suggested that intelligence be defined as [3]:
" . . . the ability of a system to act appropriately in an uncertain environment, where appropriate action is that which increases the probability of success, and success is the achievement of behavioral subgoals that support the system's ultimate goal."
Here again, the capacity for appropriate action depends on the task, and the definition of success depends on the environment.

Newell's definition of intelligence suggests a human level of cognition. Dimensions of intelligence can be attributed to a variety of systems, regardless of their complexity. Systems that exhibit very specialized and narrow capabilities that are not comparable to those of a human may still be considered to have some form of intelligence.

In this discussion, emphasize ***autonomy as a key feature of the intelligent systems*** under consideration. We want to focus on measurement of engineering parameters that enable behavior comparable to that demonstrated by living creatures. For example, we will address methods for measuring the ability of intelligent systems to:
-- sense the environment and the internal state of the machine
-- improve the performance by "bettering" the model of "self"
-- perceive and recognize objects, events, and situations and develop representation
-- remember, understand, and reason about what is perceived as well as about the actions required
-- attend to what is important and ignore what is irrelevant
-- predict what will probably happen in the future under a variety of assumptions
-- evaluate what is perceived and predicted
-- make decisions, plan, and act so as to achieve goals
-- learn from experience and from instructions

Each of these abilities can be measured in terms of accuracy, speed, efficiency, and cost/benefit ratio. For example, accuracy can be measured in terms of variance between goals and achievements. Speed can be evaluated in terms of bandwidth and latency of response. Cost can be measured in terms of resources consumed, risk, and consequences of failure. Plans can be evaluated in terms of their probability of coinciding with reality and the risk of being deviated from. Benefit can be measured in terms of payoff for goals achieved. Efficiency can be associated with computational complexity and the entropy of systems. Where appropriate, human or biological performance can provide a metric against which measurements can be made.

Specifically, we wish to discuss ideas for quantitative engineering approaches to measuring intelligence and evaluate performance of autonomous intelligent systems. Performance tests and competitions are in this class. Several types of robot and other intelligent systems competitions exist or have been proposed. They range from robotic soccer to scoring a robot's performance in war games. The machines developed to compete in these competitions tend to be narrowly focussed in their abilities and may have optimized mechanical designs to compensate for lack of "intellectual" skills.

Nevertheless, it may be possible to refine and focus these ideas to the point of having useful metrics to measure general advances in the engineering of intelligent systems.

Many additional questions arise:

- Must performance tests be domain specific?
- How does the performance depend on the architecture of intelligent system?
- Do we need calibrated facilities for quantitative evaluations?
- Would standardized test data with ground truth be useful?
- What kind of data should be collected?
- Who would perform this service?
- Can there be simple standardized tests that would be useful to individual researchers?
- Can subsystems and components be tested in isolation?

Many other issues need to be addressed (see references [4-9]). How does one compare systems of fundamentally different design? For example, how can one compare systems based on neural nets or genetic algorithms with those based on expert systems or means-ends analysis? How should systems based on reactive behaviors be compared with those based on hierarchical planning? To what extent can machine performance be compared against human capabilities? What kinds of benchmark tests and competitive contests can be designed? Is there a set of these that can be agreed upon and formalized? What should be the criteria for evaluation? These are the type of questions that we would like to be the key issues of this discussion.

References:

1. Turing, A. (1950) "Computing Machinery and Intelligence." Mind 59, pp. 433-460. Reprinted in Feigenbaum and Feldman, Computers and Thought, McGraw-Hill, New York 1963

2. .Newell, A. (1982), The Knowledge Level. Artificial Intelligence, 18(1), 87-127; Newell, A., and Simon, H. (1963), GPS: A Program that Simulates Human Thought," In Computers and Thought, Ed., Feigenbaum and Feldman, McGraw-Hill, New York.

3. .Albus, J.S., (1991) "Outline for a Theory of Intelligence," IEEE Transactions on Systems, Man and Cybernetics, Vol. 21, No. 3, pgs. 473-509, May/June

4. T. B. Gold, et al, “A Utility Approach in Multi-Agent Coordination”, Proc.. of the 2000 IEEE ICRA, San Francisco, CA 2000, pp. 2052-2057

5. E. Coste-Maniere, R. Simmons, “Architecture, the Backbone of Robotic System”, Proc.. of the 2000 IEEE ICRA, San Francisco, CA 2000, pp. 67-72

6. Q. Lin, J. W. Burdick, “On Well Defined Kinematic Metric Functions”, Proc.. of the 2000 IEEE ICRA, San Francisco, CA 2000, pp. 170-177

7. Y. von Collani, et al, “A General Learning Approach to Multisensor Based Control Using Statistic Indices”, ”, Proc.. of the 2000 IEEE ICRA, San Francisco, CA 2000, pp. 3221-3227

8. S. Schaal, et al, “Real Time Robot Learning With Locally Weighted Statistical Learning”, ”, Proc.. of the 2000 IEEE ICRA, San Francisco, CA 2000, pp. 288-293

9. K. F. MacDorman, “Responding to Affordances: Learning and Projecting a Sensorymotor Mapping”, ”, Proc.. of the 2000 IEEE ICRA, San Francisco, CA 2000, pp. 2052-2057