

PART III
SUMMARY OF PLENARY DISCUSSIONS

Compendium of the Minutes Of Plenary Discussion

Panelists: T. Balch, K. Bellman, M. Cotsaftis, P. Davis, W. J. Davis, R. Fakory, R. Finkelstein, E. Grant, J. Hernandez-Orallo, C. Joslyn, L. Reeker, E. Messina, A. Meystel, R. Murphy, C. Peterson, L. S. Phoha, Pouchard, T. Samad, A. Sanderson, A. Schultz, W. C. Stirling, G. Sukhatme, S. Wallace, A. Wild, J. Weng, T. Whalen

These notes follow the order of the papers presentation at the Workshop. Their significance is linked to the ideas and generalizations that were noticed and recorded by the panelists. Themes of these notes follow the concepts reflected in the session titles of the Workshop. Different themes generated notes of different depth and originality. Of course, this is a result of papers presented, attendees, and how stimulating the discussion was at the end of the session.

Theme 1. Features of the Industrial Intelligent Systems

The nature and embodiment of machine intelligence were discussed based upon papers:

- (1) on the description of NIST ATP-funded technology development and demonstration project
- (2) on the use of SOAR and CLIPS architectures to solve Towers of Hanoi and Quake II problems
- (3) on the definition of Task Oriented System Intelligence

The challenge was to find the consensus among these three very different aspects of the overall problem of distinguishing salient features of industrial intelligent systems. The following statements of consensus were recorded:

1. An intelligent system was initially defined as one that works to achieve goals and to survive. (The separation of the goals belonging to different time horizon is obvious).
2. The specifics of the present situation in the area of intelligent systems is in the fact that we strive to measure system *effectiveness* and *efficiency*, not *intelligence* (primarily, because we do not have the ability to meaningfully define and measure intelligence). Also, we are under the impression that we are capable of determining the effectiveness of a system. (Questions are not usually asked about the time horizon and the scope of attention in which the effectiveness and/or efficiency are evaluated).
3. Discussion was conducted on whether intelligence is inherently embodied in hardware, not software. The consensus was reached that “Hardware is one constraint on the range of a system’s admissible tasks.” (A dissenting point of view: hardware is not just a constraint but rather a carrier of intelligence, or the knowledge that is required for functioning of the intelligence).

4. An intelligent system was defined as the one that exhibits flexibility, generalization, and innovations, as limited by availability of information and ability of applied algorithms.

Theme 2: Metrics and Comparison of Alternatives: Case Studies

Paper 1: Rule-based learning can be applied successfully. Rules are derived from data from simulation or experimentation with participation of a human expert. Frameworks for knowledge-based controllers provide useful platforms for alternatives comparison.

Paper 2: Knowledge extraction from raw data can be done by using visualization with a human participating. The activities of the human can be learned by the intelligent system. Using the human-computer dialog and the visual-verbal approach allows us to extract data, properties, and models. Thus, visualization can be used for intelligence testing.

Paper 3: Intelligence can be understood as the ability to make the appropriate choices, or decisions (e.g. for robots). Intelligence makes the process of choosing simpler because of the structure that is imposed upon the decision making processes. Learning can be understood as the ability to adapt to environment (i.e. at different time scales, we will have different learning processes). Use analytic hierarchy process to define weights for IQ for robots.

Paper 4: Intelligence must be measured by looking at several abilities of the system; these abilities can be integrated by using the Additive Evaluation Method for simulating the absent metric (intelligence). One of them is based upon the idea of barter exchange and boils down to transforming all evaluations to a dollar-value. Ultimately, only the human has the best sense of each value of the intelligent function.

All papers of this Theme have something in common: the human must be kept in the loop

- (a) to measure intelligence, and
- (b) to use metric for improving control, for analysis of tools, etc.

Also, intelligence metric is still subjective and the cost-function requires human participation for evaluating the variables and assigning the weights.

Theme 3: Measuring Performance

The following salient issues were formulated:

1. Although Life and Intelligence have many similarities, they are intrinsically different in their evaluation: we can tell what is *alive* from what is *dead*. It is much more difficult to distinguish what is *intelligent* from what is *not intelligent*.

2. Formal Requirements Specifications are needed for any system at hand. Both performance and intelligence should be evaluated against the known set of their specifications. (“Ask not how much our computers can do for us, ask what we want them to do!” and “If you can specify intelligence, I can implement it!”)
3. Some of the features of intelligent systems are frequently omitted at the present time. Among them, the following should be taken into account in all cases:
 - a) disambiguation,
 - b) self-verification, and
 - c) automated synthesis (including self-synthesis)
4. The recommended approach to the system evaluation should combine the constraint-based specification with taking into account the temporal dynamic behaviors: timed vs non-timed automata. But the hard problem is the specification: it is a part of determining the dynamic behavior, too
5. Focusing upon performance measurements can be deceptive. Indeed, the system with the best performance need not be the most intelligent. If the best performance can be pre-programmed, the effort of arriving at the system with intelligence is excessive. We need intelligence *only* if the best performance is not available otherwise.
6. Thus, measuring performance without measuring the level of intelligence is not sufficient. Focus on information measures for intelligence is required, as distinct from performance. This is why the standardized tests of performance do not say anything about future functioning of a system as an *intelligent system*.
7. It would be desirable to find a simple measure of intelligence. One of the suggested measures is: *intelligence is inversely proportional to the minimum length of description for the tasks performed by a system*.
8. On the other hand, the proper functioning of the intelligent system requires satisfaction of the optimum conditions for the subsystems that support the system of intelligence, for example: *minimize total representation size*:

$$R = [R_m + R_a + R_r]$$

R_m – model representation
 R_a – representation of the algorithm
 R_r – representation of the residual part of the system
9. An example was discussed of a well described system that confirms the above projections: dexterous manipulation with multi-fingered robotic hand

10. The concept of minimizing the state description can be seen from the known importance of distinction between explicit (iconic) and implicit (abstract) state representations
11. Evaluation of the degree of automation is one of the factors that can help us in formalizing the way of evaluating intelligence. Successive technology generations characterized by increasing automation from a non-automated system to the autonomous system.
12. An opinion was presented that autonomy can be defined as the ability of a system to react appropriately to unforeseen situations [following its own determination of how to react]. Thus, the intelligent autonomy will be a subset of autonomy cases that leads to a success.
13. Nevertheless, one can demand for autonomous and semi-autonomous systems being evaluated on the scale (continuous) of the degree of autonomy observed.
14. Black Box Metrics was suggested in the White Paper. According to it the output Vector of performance was considered varied by the input Vector of Intelligence. Actually, that are other factors that affect the output of the black box:
 - a) the number of Human Operators of Complex System
 - b) the number of Loops/Operator
 - c) Size of Operational Space Automated
15. A transparent Glass Box Metrics can be introduced that allows for taking into account not only the input and output but also what is going on within the box including:
 - Richness of models implemented, e.g. using Multi-models
 - Efficiency of applied algorithms, e.g. using Anytime Algorithms
 - Sophistication of planning algorithms, e.g. using Dynamic Resource Allocation
16. Among the productive examples that can be recommended for exploring the comparative importance of the Black Box and Glass Box concept: unmanned autonomous vehicles.
17. Performance Metrics for Intelligent Systems can be analyzed by formulating “Intelligence Measuring Modules” (IMM). Their calculation is based on ordered weighted aggregation operator F, and the decision maker. The basic IMM is a set $\langle A_1, A_2, \dots, A_n : Q \rangle$ where
 - A_i are relevant measurable attributes, or features of the system
 - Q are linguistic quantifiers (such as “Most”, “At Least,” etc.)
 - $F_{w(Q)}(A, \dots, A_n) = \sum w_j b_j$
 - B_j is j th best (largest) of available A_j
18. A more general metric incorporates importance factors for A_j 's $\langle A_1, \dots, A_n : M : Q \rangle$

Theme 4: Modeling and Measuring Machine Intelligence

The common issue of the papers related to this Theme was to observe a “scorecard” or multiple capabilities and behaviors as characteristics of a single or multi-unit system and of the task environment are varied.

The attention was drawn to exploratory design for a community of what’s important over a suite of problems. Rather than conventional sensitivity analysis varying on robot or task parameter one at a time, visualization is used to enable the researcher to discover which combination of variables matter in which circumstances.

The concept of neuromorphic architectures is concerned with systems that mimic brain architecture to implement action perceptual systems which focus their attention in a closed loop interaction with the environment, an essential feature of intelligence. Behavior of such systems is systematically studied as 2 compared with that or system with “software lesions” to see if the effect of deactivating part of the simulated brain parallels the effect of lesion to the corresponding part of a real brain.

The “metric” nature of comparison can be seen in the brain organization. Survival of the organism is too slow and admits too many alternate solutions to do the job. A possible mechanism may be task completion and minimum energy matrices driving competition between incipient sets of connections during ontogeny and learning.

A program of systematic observation and development of a robot can be a part of a natural history museum, designed to be a rich social participant in interaction with humans.

In addition to a systematic qualitative research program, quantitative metrics included who followed the robot to look at aquatic dinosaurs, how long they stayed with the robot, and how well they performed on a quiz compared with these who had not interacted with the robot.

The chief bottleneck to ride application of mobile robots is not computational speed, but interactive capability, such as vision and social intelligence.

These issues can be visualized at a different resolution level. The following four languages are underlying human behavior: DNA, brain mechanisms, natural language, and written and spoken language. Research in bio-informatics is a powerful tool for understanding the lower levels in order to achieve the goal of computing with words, in which words are the input, words are the output, and the intermediate computing remains in the background.

Some highlights can be stated as follows:

- Biological systems are inspiring, they encompass the richness of the evolutionary process (the primary research already performed by Nature).
- Simulation is increasingly feasible: as the complexity of mathematical models is growing, the futility of analytical approaches gets more explicit.
- This leads to the situation when controlled experiments are easier to conduct.
- One should not expect that all types of simulation are feasible. An apprehension was expressed concerning proposals to simulate the modular-brain concept.
- One should be very careful about metaphors used in the present days terminology. One example: Social robots are different from software agents! Awareness and expressiveness make robots social, and this can be measured or estimated. This is not the case with software agents.
- *Apparent* intelligence matters; it depends on “socialness.” It is not clear presently how to judge upon this feature.
- An example can be suggested for the “socialness” evaluation: A Robot Serving as the Museum Guide.
- Man-machine interactions are primary issues in robots with “socialness” feature.
- The following list contains other factors that affect the ability to measure the level of intelligence: DNA as a part of genetic algorithms observed and/or applied, symbolic representation laws applied at each level of resolution, speech and culture of intelligent systems, physical expressions and codes of communicating intelligent systems.
- Need much more work on natural-language (NL) interfaces and computing. Eventually, the NL issues might be the key into evaluation of intelligence.

Recommendation:

1. One should systematically observe the “scorecard” of quantitative and qualitative measures of performance as one varies the capabilities of a single or multiple robot system and confronts it with a rich suite of environmental challenges (for example, groups of adults and children visiting a museum with a Robot Guide.)
2. The camouflaging of the description of processes of intelligence by gratuitous use of scientific and computational phraseology should be avoided. Let words and actions speak – keep special terminology in the background!

Theme 5: Evaluating Factors of Intelligence in Systems

The highlights of this discussion:

- Intelligence is gradual (continuous function of the features of interest) and multi-dimensional (depends of many variable factors-coordinates).

- It is preferable to assign a numerical value depending on a variable than to rank in a list that hides the dependence on the particular variables.
- The difficulty of tasks can and should be precisely measured. Thus, the evaluation of performance and intelligence might depend on prior evaluation of the “objective” complexity of tasks.
- Information-theoretical tools are especially useful for presenting the results of evaluating performance, intelligence, and the complexity of tasks..
- Factors should consider incomplete, contradictory and partially wrong information handled by intelligent systems.
- Different types of reasoning are the inherent part of the system of intelligence.
- The need for self-structuring/self-organization demonstrates itself as a component of normal learning process of the system of intelligence.
- As the process of learning develops, the system improves its own efficiency by generalizing upon similarity among multiple units of information. New, lower resolution objects emerge as a result of generalization. As this process evolves, different levels of granularity form multiresolutional hierarchies of representation.
- Standard techniques from behavioral sciences (psychology, psychometrics), biology, ecology are very useful (ANOVA, dependency analysis).
- Quantitative measures turn out to be better for efficiency of computations than qualitative/discrete ones.
- Large number of experiments are needed for Intelligent Systems if the high variance of results does not allow for forming a reliable rule.
- Sharing the results of multiple experiments is crucial for increasing the group efficiency of intelligent systems (a website and/or repository would facilitate the sharing).
- Measurement and experimentation do not provide the fully reliable value of certainty but give useful information that helps statistically the overall population of intelligent systems.
- Thus, social behavior is fundamental: it compensates for the lack of perfection of the individual intelligent system.
- Agents in a group are not totally identical, we have to find how to evaluate the optimum diversity of characteristics in the group of agents.

- There are many useful results in the intuitive approaches of the past, such as sociology, ecology, but they should be combined with contemporary information-theoretical, statistical, clustering techniques.
- Penalty-reward approach of reinforcement learning is useful for training systems as well as for measuring them without the exactly predetermined goal.
- Behavioral definitions of intelligence (Albus) can and should be put in correspondence with feature-based metrics of intelligence.
- More simple systems may behave more properly or even more “intelligently” for particular success criteria or particular environments.

Theme 6: Measuring Intelligence of Multiagent and Autonomous Networks

The major challenge for this group of intelligent systems is dealing with complexity, in particular, with exponential complexity typical for many practical cases.

Approaches:

1. Using biologically inspired systems
2. Extrasensory intelligence permissiveness
3. Metrics for embedded collaborative intelligent systems that are based on:
 - Graphical Assessment Tools
 - Various “orders” of Intelligence
 - Both applications pull and tech push
4. Domain independent measures
5. Negotiation mechanisms and coordination protocols

Theme 7: Measuring Intelligence of Distributed Systems

Four papers were presented containing a treatment of intelligent distributed systems. The following issues were highlighted:

- There is a need for highly reliable systems capable of dealing with extremely complex situations (like air traffic control...)
- These systems are typically formed of subsystems that perform specific tasks that solve some larger problem/task/or control
 - The process of decomposition is one of the key issues of analysis. An understanding should be achieved concerning the following issues: what is the principle of decomposition, how it is performed in the cases of spatial, temporal, functional, and other special cases. The possibility should be verified to aggregate the decomposed system.
 - Functional aggregation of the subsystems is a separate issue because the problem of coordination emerges which should be a part of behavior generation.

- As a result of decomposition/aggregation, the problem of intelligent control can evolve: usually, it is required to modify the actions and translate these modifications into subtasks, i.e. it is required to re-optimize the system.
- The problem of optimization is resolved at the stages of planning and control. However, the system sometimes cannot implement the optimal solution. In these cases the “satisficing” contingency should be applied.
- The problem of symbol grounding has the following practical incarnation: simulating the result of planning is frequently inadequate because a lot of underrepresented information is lacking. Indeed, the Planner envisions the desirable and even probable future, but it does not affect this future: the actuators of the system that enable and activate the process do.
- Multi-resolution representation of the system should allow for evaluating the performance and intelligence at all levels of resolution.
- Multiple independent agents are different from a consolidated system with a hierarchical implementation. The rules and laws are different of applying multiresolutional methodology to multi-agent distributed systems.
- The following features are characteristic of Key Monitor Expert Systems that start from the model/role based Expert System (e.g. for Automated Monitoring):
 - Capturing Knowledge is equivalent to creation of rules; this is a difficult issue
 - Hierarchical fault tree should be carefully constructed to distinguish the branching by resolution from the branching by decision making
 - Using intelligent systems in these cases is expensive
 - It would be prudent to anticipate the human-operator resistance
 - A carefully collected information about constraints should precede the process of action selection
- The system needs supporting “Intelligent” Agents to monitor the data
- In most of the practically known cases, the intelligent system cannot capture the knowledge of experts in full detail
- Learn the optimizing strategy has limited capabilities in practice

Theme 8: Competitions: Test Beds and Metrics

1. The following observations were made:

- Test beds are good

- USAR test beds are hard to design and even harder to design performance metrics because they are so *multifaceted*
 - finding victims/perception=>victims found
 - Interface => bandwidth used (AI is not limited to full autonomy)
 - Navigation=>coverage

- Performance based metrics which take into account the number of robots collaborating (P/N) penalize multiple robots systems (*except when Allah runs the competition*)
 - Tasks factor into this, e.g., 2 robots needed to pick up heavy box

Some other non-performance metrics are costs (monetary, energy consumption, etc.) and meeting constraints during execution (e.g., formation control)

2. The following unanswered questions were detected:

- What are the metrics for mixed initiative/adjustable autonomy vs. full autonomy? [Including HCI, adaptation to drop outs]
- Does P/N really discriminate against multiple robots in *all* tasks?
 - Can we compare intelligence versus cost?
 - How do we factor control strategy?
- Are competitions inherently flawed because they don't have the right scale/scope?

Do we have any metrics/taxonomy for task complexity?

Theme 9: Measuring Intelligence of Systems with Autonomy and Mobility

Papers stressed metrics of utility, which were argued to be more useful to designers than abstract intelligence. Two task-based metrics were combined into one task determined for the process of navigation.

The architectures discussed were constructed for different goals and applications. The system developer can only evaluate a system based on his or her own goal.

Some papers were focused on the issue of graph-based searching algorithms. The goal is to optimize the creation of the graph based on the computation resource limit.

An analysis was presented, based on their work on mental development, that a fundamental criterion is not really what a machine can do in a special setting, but its capability of developing mental skills.

The works presented in this session represent well the current status of the field: there are three areas:

1. Those that address system problems: construct a system to perform some challenging tasks. Works in this category tend to use task-specific criteria. It is not always the case that the same criteria can be used for other applications, as the presenter argued.
2. Those that address a tool that can be used for many different systems. Those tools cannot be directly used in the system until a designer has done a mapping from a practical problem that he wants to solve to the tool. This kind of work concentrates on an abstraction of a particular tool from a class of problems and thus it studies an abstract tool.
3. Another direction of the work, represented by the last presentation, addressed the automation of the developmental process.

In area 1 the human is in the loop of system design, and may choose a tool in area 2 in his or her design. In area 3, the human is not in the loop of task-specific programming. Instead the human designs a program that potentially can accomplish area 1 and area 2 autonomously, at the highly developed “adult” stage.

The field has a lot of work in area 1, which has achieved some limited success. The difficulties that face us in this area are very challenging. Although area 2 can provide some useful tools for area 1, the fundamental problem in area 1 is not the problem of tools, but rather something much more fundamental: systems are task specific and thus there are no uniformly acceptable criteria at the task level. You simply use different criteria to measure performance for different tasks.

Developmental paradigm in area 3 aims at a very different dimension. Its goal is to design a system that can develop autonomously, including learning to perform many different tasks, including such as tasks that the programmer does not know at the time of programming. Then, the capability of development becomes a universal capability, independent with what tasks that the system ends up learning to perform. In other words, it is the autonomous learning capability that the area 3 is measuring, not how well the system performs each task.

If a system has a powerful capability to autonomously learn, it will do well for many tasks it learns to perform, not just for a particular task. Interestingly, human intelligence does have a uniformly accepted set of tests for different age groups. This field is called psychometrics. These tests do not test what a human child can do, but rather whether the child can learn during the test. Thus, what is tested is the autonomous learning capability.

With this autonomous learning capability, the system can learn to perform various tasks, as long as the teaching process is well designed. This new dimension is motivated by human mental development from conception time through infancy to adulthood.

Another issue is whether it is necessary for an intelligent system to learn. This is a subject that was discussed during the workshop among some participants. We seemed to reach a consensus that if the tasks are static and are easy enough to directly program, one does not have to use machine learning. However, if the environment is unknown or partially unknown at the programming time, or the environment changes significantly during the task execution, then learning is a must. Fully autonomous learning is a new dimension known as development, which enables not only machine learning, but also automation of the learning process. Since this subject is very new, the power of this new research field is yet to be demonstrated.

Theme 10: Measuring Intelligence Taking into Account Linguistical, Biological and Psychological Factors

- Many interesting ideas are being proposed related to using language and psychological testing for measuring the intelligence, but they are not sufficiently fleshed out (at least, not yet).
- Natural language encompasses much that is important in intelligence, and certain aspects of natural language processing in the intelligent systems could even be indicative of the degree of intelligence (though even fairly retarded people and computer equipped intelligent machines are able to learn basic human languages).
- Some of the ideas related to Natural Language were presented in terms of the Turing test, and the Turing test is certainly a test that has something to do with intelligence. However, until now we are not sure what and how this relationship works and can be interpreted. Not surprisingly, Turing Test has been criticized from a lot of points of view, and our cautious view on using it as a technique for measuring intelligence seems to be justified.
- As far as Natural Language acquisition, it was not clear whether the proponents wanted to model language development or just measuring the stage of development; the first is very hard, as all of us who are interested in modeling. However, it is clear that mere measurement of "degree of development" may not tell much, and certainly won't help with the Turing test.
- Analysis of generalization processes by using Natural Language examples (summarization) can be considered illustrative of other algorithms of generalization working in living and computer-based creatures. It seems promising to explore similarities of linguistic and pictorial generalization, and eventually extend it toward symbolic generalization.

Theme 11: On the aspects of Projects related to Governmental Agencies

General observations

- The amount and the diversity of issues presented at the Workshop exceed the capability of a single specialist to encompass the situation: the parable about six blind sages analyzing an elephant: some see the trunk, some the tail, some the tusks of “intelligence”
- A taxonomy of natural and artificial intelligent systems should help to illuminate (but hopefully not eliminate!) these differences in perspective
- Ask the question: “how is the measure of a specific system’s intelligence actually going to be used?”
- Decompose the system into its constituent subsystems. But what if the “intelligence” is emergent at the system level?

Taxonomy of Intelligent Systems

- These are some examples of “Intelligent Systems”
 - Human
 - Dog
 - Cat
 - ...
 - Mobile robot
 - Industrial manipulator
 - Process controller
 - ...
- These are some Factors of “Intelligence:”
 - Sensing/perception
 - Planning/reasoning
 - Effecting/skills
- Interface/language
 - Need some sort of matrix of Factors vs
 - Types: Competencies? Requirements? ...?
- **Possible uses for the measure of a specific system’s intelligence...**
 - Answer the question “Can system A perform task X?”
 - Help determine where to spend R&D money
 - “Raise the bar” by establishing an “expected” level of achievement
 - Serve as an advertising bullet for an intelligent product
 - ...
- **Can a System “A” Perform the Task “X”?**

- DARPA supports the research and technology development in areas where the risk is high but the payoff would be significant. One aspect of this policy is to fund generously but abandon further support if it seems as though success is unlikely. DARPA program managers are strongly urged to show meaningful evidence of progress at yearly intervals. The evidence of success for the program on intelligence could be given by demonstrating that by using this approach the reliability factors could be increased.
- In the past, the evidence of success has usually been in the form of demonstrations of utility that are sometimes of questionable value in convincing potential service users of the technology that it has utility but tend to consume a significant fraction of the allocated funds. This program can result in developing fundamental techniques for testing that would be impossible to question and give them a voluntary interpretation.
- If one or more metrics could be devised for each existing governmental program, that are:

Directly relevant to the area being funded,
Related to the potential for a successful outcome, and
Measurable at reasonable cost,

it would be easier for DARPA management to evaluate progress and potentially increase the fraction of program funding that is devoted to improvement of technology.

- Since most of the advanced governmental programs are based on or include systems that can be said to embody or include “intelligence” as part of their design, funding support for the Workshop was a logical action to take.