

PART II
RESEARCH PAPERS

PART II

RESEARCH PAPERS

1. THE PHENOMENON OF INTELLIGENCE

- 1.1 Using the Metaphor of Intelligence
A. Wild, Motorola, USA
- 1.2 Technologies for Engineering Autonomy and Intelligence
T. Samad, Honeywell, USA
- 1.3 Theoretical Constructs for Measurement Performance and Intelligence in Intelligent Systems
L. Reeker, NIST, USA
- 1.4 Intelligence with Attitude
W. C. Stirling, R. L. Frost, Brigham Young University, UT, USA
- 1.5 What is the Value of Intelligence and How Can It Be Measured?
T. Whalen, Georgia State University, USA
- 1.6 On the Computational Measurement of Intelligence Factors
J. Hernandez-Orallo, University of València, Spain
- 1.7 On Definition of Task Oriented System Intelligence
M. Cotsaftis, LTME/ECE, France
- 1.8 Minds, MIPs, and Structural Feedback
R. Sanz, I. Lopez, University of Madrid, Spain
- 1.9 Fast Frugal and Accurate – the Mark of Intelligence: Toward Model-Based Design, Implementation, and Evaluation of Real-Time Systems
B. P. Zeigler, H. S. Sarjoughian, University of Arizona, USA
- 1.10 Applied Applications for Mimetic Synthesis: The AAMS Project Summary (The Intelligence of an Entity)
Garner, R. N. Bishop, USA

Using the Metaphor of Intelligence

A. Wild

Motorola, Phoenix, AZ 85018

ABSTRACT

Constructed system with autonomy can be considered as possessing intelligence, if intelligence is understood as a metaphor. It is useful to be aware of that, when defining desirable features for constructed systems, in areas such as reflecting the world (ontology), definition and pursuit of goals (teleology), or general human-like behavior (anthropomorphism). Modeling and simulating integrated systems exemplify the usage of multi-scale, multi-disciplinary representations, as a basis for increasing the autonomy of some specific constructed systems. Measuring the intelligence of constructed systems requires a Vector of Metrics for Intelligence. Its components will be defined by different means, such as conducting existence tests for essential capabilities, measuring the power to eliminate unnecessary exploration, competitions of hardware-compatible systems, or vote by a jury.

KEYWORDS: *constructed systems with autonomy, intelligence*

1. INTRODUCTION

The intelligence of the constructed systems with autonomy has to be understood as a useful metaphor, not to be stretched too far [1]. As beneficiaries of such systems, we are actually interested in their performance. The underlying assumption, however, is that building intelligence into the system, whatever its definition would be, would result in a generic and systematic way to improve their performance.

While it is relatively easy to imagine ways to measure performance, it is far less obvious how to measure intelligence, as we lack a crisp, generally accepted definition, be that for human beings, for other beings, or for artifacts.

The casual observer perceives manifestations of intelligence in multiple forms, and also will notice that somebody performing very intelligently in one situation may show what appears to be a lack of intelligence in another situation. This may suggest that intelligence is a local skill. On the other hand, some researchers intuitively feel that intelligence is an intrinsic capability of an entity, and engage in exploring the commonalities between different entities considered intelligent.

Pragmatically, the latter seems the most promising approach. If successful, it would provide the foundation for a methodology to construct systems with continuously

improved capabilities. To drive the progress, it is essential to establish metrics, ranking systems according to their intelligence. Note that for this purpose it is actually irrelevant whether one considers intelligence as a generic or a local property. Depending on the viewpoint, the ranking would be valid either within a specified sub-space or in general. However, general methods, if possible, would have clearly a wider impact.

2. LIMITS OF THE METAPHOR

A multitude of aspects can be considered as elements or capabilities necessary to support intelligent behavior. In some versions, the Vector of Intelligence has 25 dimensions. It is supported by a set of computational tools, with a system architecture counting 16 features, and is completed by a control and data acquisition system with supervisory authority, also featuring a number of capabilities. Many of these elements do justice to the view adopted by the Italian Renaissance and illustrated famously by Leonardo da Vinci: the man is the measure of all things. While this approach is quite effective, and may be often unavoidable, caution is in order to avoid excesses in at least three respects: our view of the world, our goal setting capabilities and our own being.

2.1 *Ontology*

The dimensions of the vector of intelligence and the supporting tools, architectural features and auxiliary subsystem should not be excessively isomorphic with our contemporary perception of the world.

A few centuries ago, we might have asked an intelligent system to recognize the four elements and their interactions, we would have argued about the phlogiston, and hoped that eventually an intelligent system will extract the quintessence of anything and everything. It should have recognized the planets and the major stars, and have had the ability to synchronize actions with favorable skies. The Euclidean geometry was a very pertinent model to simplify the description of the world, by accepting that concepts like a straight line do have a kind of existence. Likewise, all needed knowledge about gravity was that there exists an attraction force between two bodies, precisely equal to the Cavendish constant multiplied by the two masses divided by the square of the distance. This formula easily generated the laws derived

by Kepler from mountains of data and hundreds of years of observations. The depth of our understanding was made sensible (was measured ?) by this tremendous simplification.

Unfortunately, the space-time curvature of generalized relativity eliminated the paradigm of the straight line, and Newton's simple formula was unable to lead to a solution for three body interactions. Our present view is that the world does not admit a simple description.

When facing complexity, we tend to rely upon hierarchy to simplify interactions. Ideas about multi-resolution, multi-scale views imply a hierarchy. We tend to require that an intelligent system can do the same, being able to handle several hierarchy levels. Their number and their adequate utilization are candidates for intelligence metrics. Computational tools of intelligence define rules and procedures for crossing boundaries between hierarchy levels.

However common and widely accepted, the hierarchical representation of complexity is probably no more than the current model, and it seems reasonable to expect that it will be eventually replaced by a different view. This would also induce an evolution of the intelligence metrics derived from a model of the world, as it evolves historically.

As a matter of fact, the next paradigm may already take shape under our eyes: can one speak about the Internet as about a constructed system with autonomy, exhibiting intelligence ? And if yes, how would that intelligence be measured ?

2.2 Teleology

We consider the ability to generate goals as a leadership feature. Some philosophers consider this as the defining feature of any living beings.

However, humans, and other living creatures, pursue both explicit and implicit goals. They either conceptualized themselves the explicit goals, or receive the goals from higher authorities. In anyone of these situations, they may or may not exhibit intelligent behavior. A simple positive example is young James Watt, being given the goal to keep the pressure of a steam vessel constant. He did not conceive the goal himself, actually, he was pursuing rather different interests. It was not a goal with any recognizable intellectual challenges. But Watt generated a response that resonates until today, and will keep resonating, being, among other things, largely responsible for this workshop.

2.3 Anthropomorphism

A system scoring high on all dimensions of the Vector of Intelligence and its auxiliaries will probably pass easily the Turing test. It may do even more, it would be basically human, at least to the extent of our current understanding of the way humans are looking like. Some of the properties listed by

Neville address the ability to communicate like humans, including such things as understanding a sentence and developing knowledge. These ideas seem to relay on the perception that the more a system is similar to a human being, the more would it be perceived as intelligent.

Even if our current understanding of humans would be definitive, this approach may be an anthropomorphic trap. Actually, there is no necessity for the constructed structures with autonomy to present any isomorphism with our ideas about the human beings. Many of the most effective artifacts created by humankind are radically non-anthropomorphic, or non-biomorphic, for that matter. Starting with the wheel, radically different from a leg, yet allowing better locomotion, one can easily follow with any number of examples. A jet airplane is not a bird. A computer is not a brain. And a constructed automaton with autonomy is not a living being. There is no recognizable necessity for these artifacts to be indistinguishable from, or even similar to their closest living relatives.

If one recalls the number of words in any language describing non-intelligent behavior, one may conclude that copying too closely humans may be less than desirable.

3. PROGRESSING TOWARDS THE METAPHOR

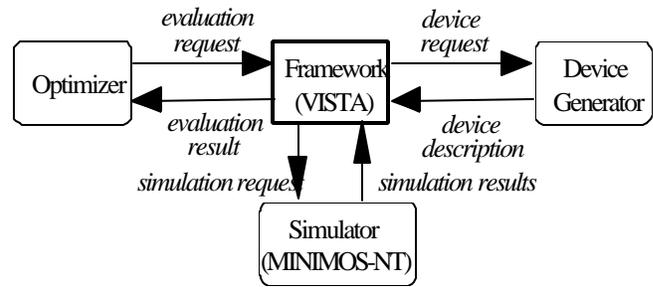
Building systems reflecting our view of the world, our purposes and our way of being, may prove productive. Multi-scale representations are probably a useful way to handle the complexity of the world in our minds, at this point in the evolution of our understanding. We can legitimately expect such representations to be useful in sciences and engineering.

The ultimate multi-discipline, multi-scale simulations are attempted by cosmologists, who hope to deduce the characteristics of the universe, 10 to 15 billion years after the Big Bang, from its characteristics when it was younger than one second.

Electronic engineers aiming to design integrated microsystems, have simpler needs: to simulate, with some quantitative accuracy, what happens on a silicon wafer within a time span from a few nanoseconds to a few hours. Microsystems are defined here as monolithic structures functionally equivalent to multi-chip systems. Increasing integration levels drive the semiconductor industry towards building system on a chip. To address this demand, design and manufacturing must integrate heterogeneous elements with traditional data processing circuits, encompassing multiple disciplines, multiple scales in space and multiple scales in time, within a coherent framework of computer aided design. Adequate modeling and simulation enables closed loop optimization and microsystem design automation.

Microsystem design must handle multi-scale modeling in time, to cope with the wide gap present in the temporal scales.

While atomistic calculations are useful for continuum simulations, molecular dynamic simulations are limited to times on the order of nanoseconds. The gap can be bridged by a meso-scale calculation, for instance using the Lattice Monte-Carlo (LMC) method to describe the hops between stable states (nanoseconds) rather than the vibration frequencies of the lattice (fractions of picoseconds). In space, multi-discipline, multi scale modeling is often required to link macroscopic reactors to microscopic integrated elements. As an example, a micromachined gear, 1 micrometer in diameter, can be analyzed using three hierarchical levels: continuum models (finite element) for the body of the wheel, molecular dynamics for gear teeth, and tight-binding for the contact between teeth. The connection is realized via a self-consistent overlap region, while keeping the time discretization in both connected domains in lock step, the whole system requiring massive parallelization at Maui Supercomputer Center.



4. MEASURING THE METAPHOR

As the Vector of Intelligence and its supporting structures are multi-dimensional, multi-faceted and quite heterogeneous, a set of metrics would probably be necessary, in the hope that if a unitary definition of intelligence would emerge, a composed metric may be put forward. The four approaches presented below are the beginning of the Vector of Metrics for Intelligence.

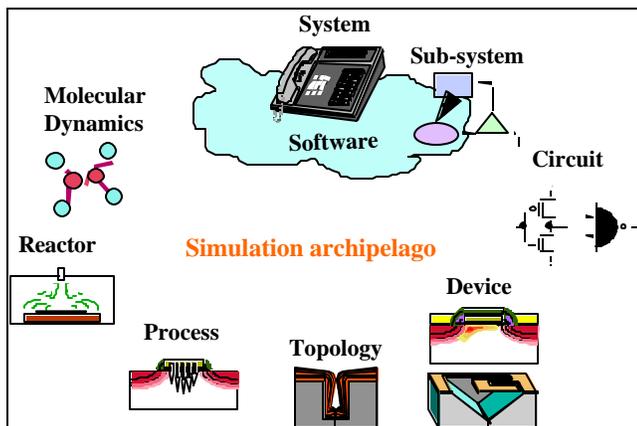
4.1 Counting features

Some features of the Vector of Intelligence and the supporting structures can be tested by a go/no go test, they either exist within a given system, or they do not. Furthermore, some of them have clear numerical definitions and can be determined by counting. The result of counting is final, as long as the structure does not evolve, or represent just an assessment at that point in time, if the system can evolve. The only open problem is how to aggregate the different dimensions of the Vector of Intelligence, so that ranking can be done.

4.2 How far away from enumeration ?

Testing for functional correctness of a system poses serious challenges even at the lowest levels. For example, testing the hardware of a microprocessor, a finite state machine, is conceptually easy, yet unsolvable practically. Theoretically, a test can run through all possible transitions between states, with all bit configurations at the external inputs, comparing at each step the outputs with the specification. The number of states and transitions is finite, yet so large, that the test of a 32 bit processor running at 1GHz would take a time longer than the age of the Universe.

To reduce the number of tests, one can use additional switching elements to reconfigure the structure to a finite state machine of lower complexity. If the logic gates and storage elements in the finite state machine have been defined algorithmically, one can safely accept that the functionality would be correct, if no physical defects are present. In this



Currently, the multiple disciplines involved in microsystems are either unconnected, building an archipelago, or put together by human programmers in an ad-hoc manner. Active research, however, is aimed at systems able to build bridges between the isolated domains, as a pre-requisite for using optimizers in closed loop. This technique allows the correlation between decisions at one manufacturing step and the system level features and performance.

Using an optimizer at the meta-level to manage the design process brings the system one step further. Many features would be required to incorporate these or similar functions in a constructed system with autonomy, exhibiting some intelligence.

This “bottom up” progression towards a development system with autonomy increasingly adds features included among the dimensions of the Vector of Intelligence. This seems a promising way towards the next challenges in engineering, believed to be nanosciences, biological systems, and last but not least, robotics. Searching for their intelligent features would surely provide underlying commonalities and accelerate the progress.

case, the simplified structure may be used to proof that all the desired logic gates and storage elements (a few 10 or 100 million of them on contemporary chips) are present, functional, and properly connected. These methods, currently used, are still unable to provide satisfactory test coverage. At a more abstract level, formal analysis of the structures is researched as the next opportunity to achieve it. If one adds to the testing the requirement to proof that a system or a piece of software is providing optimum responses in all cases, the complexity of the task is inhibiting.

In general, a measure of intelligence could be how much of the space to be investigated is not explored through enumeration.

This is almost isomorphic with some areas of scientific knowledge. For instance, the postulates of thermodynamics, to be accepted rather than demonstrated, point out what is impossible to achieve, saving us huge efforts, like trying to build all possible cases of perpetuum mobile of the first and second species, in addition to trying to reach absolute zero. Obviously, the postulates are very effective in eliminating an infinity of pointless attempts.

4.3 *Contests*

Intelligent systems are expected to perform well in uncertain situations, and direct competition among systems might be an appropriate way to generate uncertainty, providing means to rank them.

Examples of competitions are robot wars, fire-fighting robot contests, or robot-soccer tournaments. It is necessary to define the contests such that they address either the body or the mind of the systems in competition. Robot wars address obviously both. Athletic capabilities, rather than intelligence, also determined the outcome of the last World Cup for Robot Soccer, at which one team had access to more powerful motors than the other teams.

To dissociate the two components, an easy way would be to organize games between robots mechanically identical, but driven by different minds, a luxury seldom available with human beings.

4.4 *Vote*

Capturing all elements necessary for intelligent behavior is a complex and controversial endeavor. The Vector of Intelligence and supporting features, even after unnecessary anthropomorphic features have been eliminated, still has dimensions judged by perception.

Contemplating the behavior of living beings, one would readily identify some that would be spontaneously perceived as non-intelligent (stupid), while a whole range would be rather neutral, neither intelligent nor stupid. An alternative approach to building intelligent systems, could be to address the topic of

building non-stupid systems, specifying what they should NOT do.

For instance, they should not persist in error. A non-stupid system would recognize a hopeless situation, and change its behavior or method. This distinguishes intelligence from blind instinct: ants keep building their houses even after the eggs have been removed. Although methods have been defined and implemented for quite some time to avoid stalling, quite sophisticated autonomous systems on a remote Planet still got stuck, as do soccer playing robots. When a player manages to get unstuck by spinning, the human observers cheer. However, the opposite result is achieved, when players start spinning without a recognizable reason.

Given the subjective component in characterizing behavior as being intelligent, one could also envision scoring by the vote of a human jury. This would be similar to the methods used in some sports such as skating, in which a jury gives two notes: one for the technical merit, one for the artistic impression. After all, contests and games are entertainment, and audiences are entitled to have some fun.

5. REFERENCES

[1] White Paper of the Workshop on Performance Metrics for Intelligent Systems,
http://www.isd.mel.gov/conferences/performance_metrics

Technologies for Engineering Autonomy and Intelligence

Tariq Samad

Honeywell Technology Center
3660 Technology Drive
Minneapolis, MN 55418, U.S.A.
tariq.samad@honeywell.com

ABSTRACT

A critical need for a high performance autonomous system is the ability to generate appropriate responses when faced with conditions that were not explicitly considered during off-line design. This paper emphasizes three technical concepts as essential for meeting this need: multimodels, anytime algorithms, and dynamic resource allocation. An example from ongoing research in the autonomous uninhabited aerial vehicle domain is used to illustrate the concepts. Some competing concepts are discussed, and connections with consciousness and metrics are outlined.

Keywords: *Autonomous systems, multimodels, anytime algorithms, resource allocation, uninhabited air vehicles, consciousness.*

1. INTRODUCTION

Society, industry, and government are all exhibiting increasing interest in autonomous and semi-autonomous systems—complex engineered artifacts that require minimal or no human involvement for their operation. The motivations for this interest range from cost-efficiency to environmental safety to national defense. Potential applications are everywhere, especially where human operation is infeasible or dangerous: warfare, deep space missions, terrorism countermeasures, and toxic material handling are examples that come readily to mind.

From one perspective, it could be argued that the history of automation is the history of progress in engineering autonomy. We have been successful in automating ever-higher levels of operation, from regulatory control to supervisory control on upward. The Wright Flyer required the human pilot to perform the inner-loop control function. Today's commercial aircraft can fly from point A to point B, automatically closing the loop on not just the inner loop but also outer loop, handling qualities, and waypoint following functions.

But autonomy is much more than automation. Today's engineered systems may be highly automated, but they are brittle and capable of "hands-off" operation only under more-or-less nominal conditions. As long as the system only encounters situations that were explicitly considered during the design of its operational logic, the human element is dispensable. As soon as any abnormal situation arises, control reverts to the human.

An autonomous agent must be capable of responding appropriately to *unforeseen* situations—that is, situations unforeseen by its designers. Some degree of circumscription of a system's operating space will always exist, since survival under every environmental extreme is inconceivable, but "precompiled" behaviors and strategies are not sufficient for effective autonomy.

Below, I first discuss some features and characteristics that I believe are necessary for engineering high-performing autonomous systems. Next, in Section 3, an example from work in progress—which is focusing on the development of autonomous capabilities for uninhabited aerial vehicles—is presented. Section 4 discusses some alternative perspectives on engineering autonomy, followed by a selective review of the consciousness controversy. I conclude with a measurement-related note.

Parts of this paper are adapted from (Samad and Weyrauch, 2000) wherein some further elaboration can be found.

2. ASPECTS OF AUTONOMY

What does it mean to be able to react appropriately to unforeseen situations? To be capable of exhibiting behaviors that are not precompiled? I would like to emphasize three technical concepts: multimodels, anytime algorithms, and dynamic resource allocation. These are discussed below, and a brief digression on the topic of hierarchy is also included.

2.1 *Multimodels: Explicit representations of heterogeneous knowledge*

In the absence of a sufficiently rich built-in library of canned responses to specific situations, an agent must be able to rely on an explicit, algorithmically manipulable knowledge base. Instead of reflexive responses being built in, the knowledge base required to generate responses deliberately must be incorporated.

The knowledge base must capture relevant details about the capabilities of the autonomous agent, its environment, other agents it expects to be interacting with, its tasks or objectives, etc. These "models" need not be perfect; they represent what the agent believes, not objective truths. But, almost regardless of their fidelity, they allow the agent to reason and to determine responses to a potentially hostile world. The effectiveness of the responses will be a function of the fidelity of the models (in part), but, I would maintain,

autonomy and effectiveness are separable. Stupid intelligence is an oxymoron; stupid autonomy is not. (In most of this paper, however, I do not make a careful distinction between intelligence and autonomy.)

I use the term multimodels to refer to multiple, heterogeneous knowledge representations. We later discuss a domain-specific example, but here I would like to note one property of multimodels that is likely to be useful across domains. The degree of precision and accuracy of knowledge that an autonomous agent must consider will vary with the situation it finds itself in. In some cases, disparate models may be used to capture different levels of detail. However, a greatly preferable option is a unified modeling framework that is capable of providing estimates or predictions at multiple levels of resolution, the level in effect at any time being specifiable by a higher level function.

2.2 Dynamic resource allocation and anytime algorithms

An autonomous agent must be able to dynamically manage its processing and other (sensing, actuation, communication, power) resources. In the face of multiple competing demands and objectives, each of which requires individual algorithmic attention, an agent cannot generally afford to examine any exhaustively. The world does not wait for closure of contemplation.

Thus, tradeoffs must be made in real-time, to decide how inevitably inadequate resources must be apportioned to the multiple demands on them. This is an issue that generally gets little attention from the intelligent systems community, yet it is no less critical than the issue of designing algorithms for information processing for autonomous systems.

Different processing tasks have different criticalities, deadlines, and other properties. Some tasks may need to be executed on a fixed periodic basis, others may be event-driven, others yet may be continually ongoing. This variety is suggestive of the complexity of real-time resource management for autonomous systems.

Of particular interest for autonomous operation are “anytime” algorithms—algorithms that are able to flexibly exploit available computational resources. Beyond a certain minimum execution time that it may require to generate an initial candidate solution, an anytime algorithm can iteratively improve on this solution over time. Randomized algorithms such as evolutionary computing are prototypical examples.

Resource management in current control systems presents an illuminating contrast with the needs for autonomous operation noted above. All control systems today have to address resource constraints. This is done by determining ahead of time—during the design process—precisely which

tasks will need to be executed under what conditions. Task execution schedules can then be precomputed and defined. This static scheduling approach is infeasible for autonomous systems.

2.3 Hierarchies, but not strict ones

The sophisticated information processing systems we currently engineer are almost always hierarchical. Further, the design methodology that is proposed in today’s technoculture emphasizes strict, hierarchically structured processes. Hierarchy as an engineering design heuristic has much to recommend it, but I would assert that it is a mistake to assume that all intelligent systems must be analyzable as strictly hierarchical. One need only look at the central nervous system of any organism one thinks of as intelligent (e.g., the human brain) as evidence. There is certainly structure to the brain, but a formal, strict hierarchy is a counterfactual insistence. Bypass connections, reflex reactions, affective conditioning, many intriguing pathologies—these are all indicative of an organization that is better thought of as a web than a tree, or at least as only loosely hierarchical.

As an example, see Figure 1. Elements of the figure resemble the typical multilayer hierarchical architectures that attempts at engineering autonomous systems often adopt (i.e., the organization as shown of the spinal column, the brainstem, the thalamus, and the cerebrum). However, additional pathways are also present, forming prominent and crucial bypass structures and feedback loops.

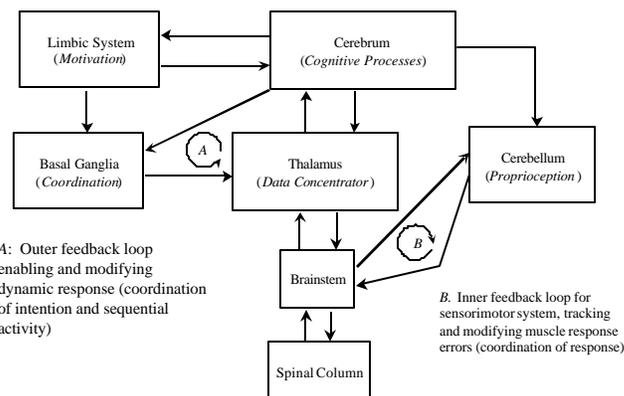


Figure 1. Simplified architecture for primate central nervous system (figure courtesy of Blaise Morton).

3. EXAMPLE: ROUTE OPTIMIZATION FOR AN UNINHABITED AUTONOMOUS VEHICLE

We briefly discuss here some ongoing research at Honeywell Technology Center, targeted toward the development of algorithms and software mechanisms for uninhabited air vehicles (UAVs), with specific emphasis on demanding military applications. Multimodels, anytime

algorithms, and dynamic resource allocation feature prominently in our research.

An example of a multimodel knowledge base for route and trajectory optimization in a UAV is shown in Figure 2. The figure shows a (wavelet-based) multiresolution time/frequency model of a trajectory. By selectively setting specific parameters—each associated with one of the boxes in the top graphic—to zero, the space of trajectories can automatically be constrained so that different segments of the trajectory are defined in more or less detail as appropriate for a given situation. Trajectory optimization is then conducted over the enabled parameters, ensuring that computational resources are used efficiently. Under normal conditions, we can expect that the resolution profile would gradually decrease over the optimization horizon. The figure also shows multiresolution models of aircraft dynamics and terrain; these and other models are necessary to check various constraints on a hypothesized trajectory and to calculate the cost function for optimization. (See Godbole, Samad, and Gopal [2000] for more details.)

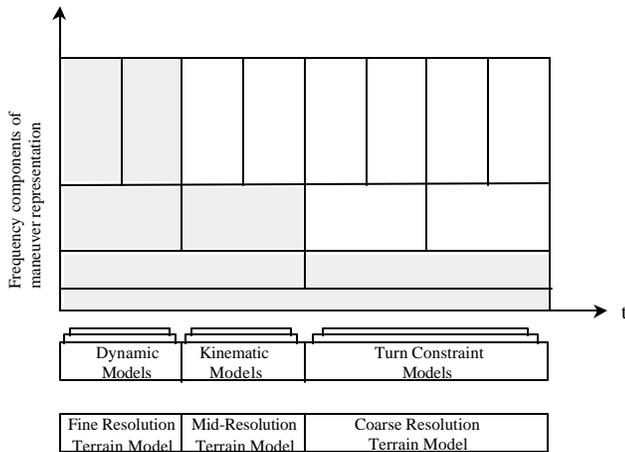


Figure 2. Multimodels for trajectory optimization for an autonomous aircraft.

This multimodel approach has been integrated with an anytime algorithm for route optimization, and a simulation result is shown in Figure 3. A UAV is skirting a threat area when a target model (including the target’s coordinates) is communicated to it. The original route (not shown in the figure) was not overflying the target area but instead adopting a low elevation radar-evading route over a ravine. Once the target is detected, the online trajectory optimization algorithm is executed. In this case, greater resolution is desired over a medium horizon interval, and minimizing the previous cost function for low flight is considered less important than rapidly generating an alternative route that overflies the target area. As the UAV continues its flight, incremental re-optimizations are performed at regular intervals, with the computational resources expended on these optimizations varying

continuously depending on the particular objectives and models under consideration at that time.

We currently use an evolutionary computing algorithm—an extension of the algorithm outlined in (Samad and Su, 1996)—for optimizing the trajectory. The EC algorithm searches over the space of nonzero coefficients in the multiresolution wavelet-based representation noted earlier.

As I hope this example illustrates, the concepts of multimodels, anytime algorithms, and dynamic resource management are related in that effective autonomy requires the integration of all of them. Given a particular situation that requires an autonomous agent to react, it must be able:

- to access the knowledge it has that is relevant to the situation in the context of its goals and abilities;
- to flexibly reason about its decision and control options, adapting the level of scale and resolution in its processing to the situation and objectives;
- to tradeoff competing demands and requirements in the face of resource limitations.

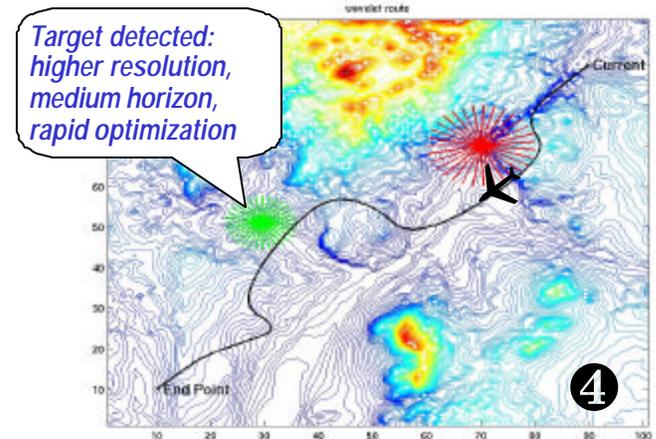


Figure 3. A frame from a simulation example of active multimodel control for trajectory optimization.

4. ALTERNATIVE PERSPECTIVES

There are, however, other reasonable solutions and perspectives to engineering autonomy that are being proposed, and a few are briefly noted in this section.

4.1 Model-free autonomy

It seems reasonable to correlate the autonomy of a system with the fidelity or scope of the models accessible to it, a connection I have made above. The richer the explicit

representations of its environment, itself, its collaborators, etc., that a system contains (regardless of whether these representations are acquired through learning or are hardwired by a designer) the more likely that an engineering system can operate effectively without continuous human supervision. So a model that can be symbolically manipulated may be seen as a necessary condition for autonomy.

But consider (as much research in intelligent systems is starting to do) an ant. There are certainly properties of ant behavior that we would be delighted to be able to incorporate within constructed systems with autonomy. An artificial ant, if we were able to construct one, would be considered to be a system with some non-trivial degree of autonomy.

Or, if the capabilities of an ant do not warrant the “autonomy” label, what about an ant colony? A million ants no more make an explicit, manipulable model of the world than an ant by itself.

The most prominent exemplar of this line of research in autonomous systems is the “subsumption architecture” of Brooks (1991), a central tenet of which is that the world can be its own model. No representations are needed—in fact, they are seen as harmful since in dynamic and ever-changing environments they can rapidly become outdated.

4.2 *Is biology the only model?*

Today, all the truly autonomous systems that exist are biological ones. It therefore seems appropriate to mimic salient features of biological systems in the design of engineered autonomy. However, an alternative viewpoint may lead us to question such biomimicry. Most human engineering, an endeavor that has enjoyed considerable successes, has not drawn design inspiration from biological principles—airplanes are an obvious example.

Architectural sketches of brain organization (as in Figure 1) may be dismissed as irrelevant by this argument.

Of course, until some non-biologically-inspired autonomous artifact is produced, the study of existing autonomous systems (i.e., biological ones) should be helpful. But it can legitimately be argued that biology need only be a weak model.

4.3 *Autonomy need not be physically grounded*

Our discussion above has exemplified autonomous systems with UAVs, and most research in autonomy focuses on vehicular systems (terrestrial, undersea, or in air or space). While autonomous vehicles are a particularly exciting prospect for future engineering systems, autonomy, as a property, should not be considered constrained to physically mobile platforms.

In fact, it is important to consider autonomous systems that are not vehicles, since a broader understanding of autonomy is contingent on an understanding of the full spectrum of the

topic. Different application areas will have specific characteristics. For example, in the process industries there is a continuing drive to increase the level of automation in plants, sometimes even quantified by a “number of loops per operator” metric. An autonomous decision and control system for an oil refinery will have to deal with issues related to high dimensionality (a refinery can have 20,000 sensors and actuators), significant delays due to material transport (dead times can be on the order of hours), and the lack of full state feedback.

At an even greater remove from physicality, we can contemplate autonomous computer and communication networks, which need operate only in the “virtual” realm.

5. CONSCIOUSNESS—REQUIREMENT OR RED HERRING?

The notion of developing engineered sensors or actuators, or even low-level models of computation, that are based on biologically gleaned principles is uncontroversial.

Embodying higher-level cognitive capabilities in computational systems, however, is another matter. Some would argue that the sorts of phenomena found in the brains of humans cannot even in principle be realized by the sorts of machines we are contemplating. The levels of autonomy, intelligence, and adaptability exhibited by humans are thereby excluded (the argument goes) from realization in engineered systems.

The concept of consciousness lies at the center of this controversy. I take it as given that human-like performance by a machine requires the machine to have something akin to consciousness—an ability to reason about and reflect on its own behavior, not just “blindly” follow preprogrammed instructions.

There are two theoretical limitations of formal systems that are driving much of the controversy—the issue under debate is whether humans, and perhaps other animals, are not subject to these limitations. First, we know that all digital computing machines are “Turing-equivalent”—they differ in processing speeds, implementation technology, input/output media, etc., but they are all (given unlimited memory and computing time) capable of exactly the same calculations. More importantly, there are some problems that no digital computer can solve. The best known example is the halting problem—we know that it is impossible to realize a computer program that will take as input another, arbitrary, computer program and determine whether or not the program is guaranteed to always terminate.

Second, by Gödel’s proof, we know that in any mathematical system of at least a minimal power there are truths that cannot be proven and falsehoods that cannot be disproved. The fact that we humans can demonstrate the

incompleteness of a mathematical system has led to claims that Gödel's proof does not apply to humans.

In analyzing the ongoing debate on this topic, it is clear that a number of different critiques are being made of what we can call the "computational consciousness" research program. In order of increasing "difficulty," these include the following:

- Biological information processing is entirely analog, and analog processing is qualitatively different from digital. Thus sufficiently powerful analog computers might be able to realize autonomous systems, but digitally based computation cannot. Most researchers do not believe that analog processing overcomes the limitations of digital systems; the matter has not been proven, but the Church-Turing hypothesis (roughly, that anything computable is Turing-Machine [i.e., digitally/algorithmically] computable) is generally taken as fact. A variation of this argument, directed principally at elements of the artificial intelligence and cognitive science communities, asserts that primarily symbolic, rule-based processing cannot explain human intelligent behavior.
- Analog computers can of course be made from non-biological material, so the above argument does not rule out the possibility of engineered consciousness. Assertions that the biological substrate itself is special have also been proposed. Being constructed out of this material, neural cells can undertake some form of processing that, for example, silicon-based systems cannot. Beyond an ability to implement a level of self-reflection that, per Gödel, is ruled out for Turing machines, specifics of this "form of processing" are seldom proposed, although Penrose's hypothesis that the brain exploits quantum gravitational effects is a notable exception (Penrose, 1989). (It is worth noting that no accepted model of biological processing relies on quantum-level phenomena.)
- It has also been argued that intelligence, as exhibited by animals, is essentially tied to embodiment. Disembodied computer programs running on immobile platforms and relying on keyboards, screens, and files for their inputs and outputs, are inherently incapable of robustly managing the real world. According to this view, a necessary (not necessarily sufficient) requirement for an autonomous system is that it undertakes a formative process where it is allowed to interact with the real world.
- Finally, the ultimate argument is a variation of the vitalist one, that consciousness is something extra-material. For current purposes this can be considered a refrain of the Cartesian mind/body dualist position. Modern variations on this theme include Chalmers (1995)—an article that also includes a rebuttal by Christof Koch and Francis Crick.

The issue of consciousness in machines has captured the imagination of many as a result of the famous (or notorious) Chinese room thought experiment suggested by John Searle (1980). Searle imagines himself locked inside a room, unable to communicate with anyone outside except through slips of paper passed through a slot in the door. These slips of paper are written in Chinese, a language Searle has no knowledge or understanding of. However, Searle has been given a voluminous "script" that details (in English) the algorithmic manipulations that he should carry out upon receipt of messages. Some of the messages can have questions written on them, others may describe a story. Searle allows that the script is perfect in that the manipulations result in responses that Searle can transcribe (the symbols that he reads, manipulates, and writes are meaningless squiggles to him) and pass back to his interrogator. These responses are in fact appropriate in context; to the person outside, Searle must understand Chinese. The point of the Chinese room (thought) experiment is that knowing how the responses were generated we would not say that Searle "understands" Chinese. This is a critique of one school of thought that maintains that rule-based algorithmic processing is sufficient for understanding. Variations of the experiment and the argument have since been directed at other types of automated mechanisms.

Consciousness is a multifaceted phenomenon. I would maintain that reflective, deliberative decision making is an important element, although admittedly not the only one. Thus the technical concepts discussed earlier—multimodels, anytime algorithms, dynamic resource allocation—which, I argued, are essential for high-performance autonomous behavior, are by the same token necessary correlates of consciousness. (Our observations of) our own conscious processing support(s) this contention—we dynamically allocate cognitive resources as appropriate for an unforeseen situation, scale the precision and resolution of our processing accordingly, and rely on our knowledge of the various systems and phenomena that constitute our environment.

6. TOWARD METRICS

Even for humans, testing and quantifying intelligence is a controversial activity. The difficulty of compressing the multifaceted nature of intelligence into one scalar quotient has led to proposals to consider "intelligence" not as one unitary quantity but as a collection of properties that are mutually incommensurable (e.g., Gardner, 1983).

But humans, as a species, have much in common. We all have the same sensory apparatus; the same physiology, more or less; the same innate drives; the same communication apparatus; etc. If quantifying intelligence is so problematic for humans, one can wonder whether it is even sensible for artificial systems, which may have little or nothing in common. Comparing and contrasting the

intelligence of an intelligent search engine for the Web with the intelligence of an autonomous vehicle is a challenge that is not only huge but perhaps unaddressable. We will need to decompose the notion of intelligence in this case too, except that instead of a handful of separate factors we might end up with a much larger number.

The technical concepts I have focused on in this paper can all be considered dimensions along which autonomy and/or intelligence can be measured. The extent to which an agent has available explicit models of relevant phenomena and systems, the scaling capabilities of the anytime algorithms available to it, and the sophistication of its adaptive computational resource allocation mechanisms, all bear on how well the agent will perform in a complex, dynamic world. More research is needed before these connections can be formalized or quantified—I have been concerned here with just their identification.

Acknowledgement

The research discussed in this paper is supported in part by the U.S. Defense Advanced Research Projects Agency (DARPA) under contract number F33615-98-C-1340.

References

- Brooks, R. (1991). Intelligence without representation, *Artificial Intelligence*, vol. 47, pp. 139-159.
- Chalmers, D. (1995). The puzzle of conscious experience, *Scientific American*, pp. 80-86, December.
- Gardner, H. (1983). *Frames of Mind*. New York: Basic Books.
- Godbole, D., T. Samad, and V. Gopal (2000). Active multi-model control for dynamic maneuver optimization of unmanned air vehicles. *Proc. Int. Conf. on Robotics and Automation*, San Francisco, CA.
- Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford Univ. Press.
- Samad, T. and T. Su (1996). On the optimization aspects of parametrized neurocontrol design. *IEEE Transactions on Components, Packaging, and Manufacturing Technology*. vol. 19, no. 1, pp. 27-36.
- Samad, T. and J. Weyrauch, eds. (2000). *Automation, Control and Complexity: An Integrated View*. Chichester, U.K.: John Wiley & Sons.
- Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, vol. 3, pp. 417-458.

Theoretical Constructs and Measurement of Performance and Intelligence in Intelligent Systems

Larry H. Reeker

National Institute of Standards and Technology

Gaithersburg, MD 20899

(Larry.Reeker@NIST.gov)

Abstract

This paper makes a distinction between measurement at surface and deeper levels. At the deep levels, the items measured are theoretical constructs or their attributes in scientific theories. The contention of the paper is that measurement at deeper levels gives predictions of behavior at the surface level of artifacts, rather than just comparison between the performance of artifacts, and that this predictive power is needed to develop artificial intelligence. Many theoretical constructs will overlap those in cognitive science and others will overlap ones used in different areas of computer science. Examples of other “sciences of the artificial” are given, along with several examples of where measurable constructs for intelligent systems are needed and proposals for some constructs.

Introduction

There are a number of apparent ways and certainly many more not so apparent ways to measure aspects of performance of an intelligent system. There are a variety of things to measure and metrics for doing so being proposed at this workshop, and it is important to discuss them. To develop a measure of machine intelligence that is supposed to correlate with the system’s future performance capability on a larger class of tasks considered intelligent would be analogous to human IQ. That would require agreement on one or more definitions of machine intelligence and finding a set of performance tasks that can predict the abilities required by the definition(s), and still might not say much about the nature of machine intelligence or how to improve it.

One reason that metrics of performance (and perhaps, of intelligence) are needed is that they directly address the fact that it has been difficult to compare intelligent systems with one another, or to verify claims that are made for their behaviors. Another reason is that having measurements of qualities of any sort of entity provides a concrete, operational way to *define* the entity, grounding it in more than words

alone. All of these aspects - *comparability*, *verifiability*, and *operational grounding* - were undoubtedly at least part of what Lord Kelvin meant about measurements providing a feeling that one understood a concept in science. (See the preamble to this workshop [Meystel *et al* 00]: "When you can measure what you are speaking about and express it in numbers, you know something about it.")

The measurements that form the primary topic of this paper are of a different type. They are ones that look ahead to the future, when the intelligent systems or artificial intelligence* field is more mature. The notion of mature field is defined here in terms of scientific theories that predict the performance of the systems on the basis of the underlying science. It is suggested that really valuable measurements require reliable predictions of this scientific sort, rather than just ways to compare the technological artifacts based on the science. To do this, it is necessary to develop theories containing measurable theoretical constructs, as will be discussed below.

The discussion of metrics for attributes of theoretical constructs herein does not conflict in any way with the idea of overall system measurements, comparisons, or benchmarks, which are useful for the purposes mentioned above. In fact, it is a philosophical problem to decide where theoretical constructs stop and empirical constructs begin. Measurements of artifacts will be referred to as **surface measurements**, those of a more theoretical nature as **deep measurements**, terms borrowed from Noam Chomsky’s [65] terms for levels of syntactic description. The question of “how deep” can be left open at this time. This paper advocates looking for measurable theoretical constructs at the deeper level that will predict surface behaviors at the level of the system or subsystem, or of an entire artifact.

* The latter term will be used herein because the shortened form, “AI” is more common than “IS”.

The remainder of the paper explains the form that we will expect for AI theories in the future if they are to qualify as scientific theories and suggests theoretical constructs that may have measurable properties. It will discuss existing constructs that are developing as candidates for deep metrics and how they may relate to surface measurement. It will compare them to constructs in existing scientific theories at deep and surface levels. It will suggest that they will naturally relate to constructs from the artificial and natural sciences, specifically from cognitive science and computer science.

Computation Centered and Cognition Centered Approaches to AI

At all levels, from surface to deep, the constructs to be measured may depend on the approach taken to AI. There are two distinguishable approaches that have been taken over the years, which we will call “computation centered” and “cognition centered”*. The computation centered approach focuses on how certain tasks can be accomplished by artificial systems, without any reference to how humans might do similar tasks. We do not usually think of numerical calculation as AI, but if we did, we would have to think of the way it is done as computation centered. There is no particular reason to make it cognition centered.

In the cognition-centered approach to AI, the tradition is to discover human ways of doing cognitive tasks and see how these might be done by intelligent systems. Sometimes the motivation for this approach has been to try to find plausible models for human cognitive processes (cognitive simulation), but for AI purposes, it has often been a matter of using human clues to try to accomplish the computation centered approach. Some researchers feel that developing the artifacts using cognitive ideas may lead to more robust AI systems (using “robust” in the sense that the system is not narrow or “brittle” in its intelligent capabilities). But it is a natural way to think about the developing AI capabilities, since not all areas related to intelligent activities have been

* In the email exchange leading up to the Workshop, a third approach, “Mimetic Synthesis”, whose prime concern is the “Turing test” one of representing a computer to a human user as if it were another human, was distinguished from the two mentioned by Robby Garner. It is a good distinction, though like the others, the boundaries are not always clear.

explored and reduced to mathematical methods to the extent of numerical calculations, or even of mathematical logic, which might directly facilitate a computation centered approach.

Mathematical logic makes an interesting case for pointing out that most AI researchers in practice blend the computation centered and cognition centered approaches, since it is formalized, yet still can be approached in a cognition centered way. Computers actually implement mathematical logic, which is essential in control statements of programming languages. However, actually proving theorems in logic (beyond propositional logic, where truth-table methods can be used), is a creative intelligent activity. There, things become more complex, in different ways. The first complexity is that is a *creative* activity and we do not really understand even how people do it. Secondly, it is *informationally complex*: there are inherent undecidability problems in logics of sufficient richness for most interesting purposes.

In attempts to make it easier for humans to prove theorems, natural deduction methods were invented by Gentzen [34] and developed by a number of people, notably Fitch [52]. In a sense, natural deduction can be thought of as a computation-oriented version of theorem proving, taking away some of the mental work of creativity. But this does not change the inherent informational complexity problems, which provide inherent limits on computability.

Going beyond logic to general problem solving one finds some empirical studies of effective ways in which humans do it that antedate the computer. One of them, means-ends analysis, was codified in the General Problem Solver (GPS) program of Newell and Simon. [63] (See also Ernst and Newell 65). For programs in the GPS era, it was in the spirit of that work to attempt measurement of the extent to which the program could mimic human behavior. This was done by also studying verbal protocols of people solving the problem. Any way of comparing those to the performance of the program was still pretty much a surface measurement. Such surface measures of cognitive performance, are also the heart of the Turing test [Turing 50], but do not tell us much about what is happening deeper in the system, as Joseph Weizenbaum showed with Eliza [66] (and emphasized in an ironic letter [74]). In more recent times, case-based methods have been advocated [Kolodner 88] as relating to the way some people solve problems and they do

look very promising. Some of the constructs from these problem-solving methods will be mentioned below.

Though computation centered and cognitive centered approaches blend well, the measurements that occur to the developers in the two approaches will naturally differ, and this is particularly true as one tries to go to a deeper level by using constructs that are based either on cognition or on computation. In other words, AI may have measurable constructs coming from at least two different sources, the computation side and the cognitive side. This fact has some interesting implications as one looks at the measurement of deeper constructs, which may have to be reconciled with both approaches to be meaningful.

The Structure of Scientific Theories

Today's views of scientific theory have changed from those held in the 19th Century, Lord Kelvin's time. The bare-bones version of a scientific theory today is that it consists of a model composed of abstract **theoretical constructs** and a **calculus** that manipulates these constructs in a way that can account for observations and accurately predict the value of experiments. The model is as central today as was the notion of measurement to Kelvin. The theoretical constructs have a relation with observed entities, properties and processes that may be quite abstract, not necessarily readily available to human senses, but following directly from calculations based on the theory. There are a number of principles applied to a model that give us increased confidence in the theory, but the one most relevant here is that we can measure the observed entities to confirm the predictions of the theories. So Kelvin's concern has been preserved, but augmented, in today's view of theories.

It is relevant to observe that the "calculus" mentioned above is used in the dictionary sense "a method of computation or calculation in a special notation (as of logic or symbolic logic)". That means that it may be numerical or non-numerical. In fact, as Herb Simon and Allen Newell [65] pointed out, there is no reason that the calculus cannot be expressed in the notation of a computer program, the better to speed its manipulation of the theoretical constructs.

For scientific theories in AI to be respectable, there will be certain requirements on them, and these affect whether they are accepted

or not and whether the theories in which they occur are accepted. The late Henry Margenau had a pragmatic treatment of these requirements in his book *The Nature of Physical Reality* [Margenau 50]. A working Physicist as well as a philosopher, Margenau stressed that no amount of empirical evidence was scientifically convincing by itself, since it did not specify a unique model; and he also stressed the need for the binding of theoretical constructs to one another in a "fabric". This fabric was made up of theory and of mappings to empirical data. The theory was convincing to the degree that certain criteria were met - not a "black and white" situation, but one of degree. One of the criteria was the extent to which the models and constructs were extensible to larger and larger areas of scientific endeavor. As the fabric of the theory became larger and stronger, it became more difficult to rip it asunder.

Perhaps our emphasis on finding metrics can solidify the theoretical constructs of the field, as well as providing a means of measuring progress. The key to doing this is not to think of evaluation only as measurement of some benchmarks or physical parameters ("behaviors") that are manifested in the operation of the systems being evaluated. We need to be thinking in terms of the inner workings of the systems and how the parameters within them relate to the measured externally manifested behaviors.

One of Lord Kelvin's special interests was temperature. Temperature is of course something that we experience, something not wholly abstract. Certain physical properties are related to temperature, and the most easily observed is freezing and boiling of water. It took some scientific discovery to realize that each of these phenomena always take place at a particular (with a few reservations, like altitude and purity of the water), but still, those are concrete embodiments. Temperature has been a subjective attribute during most of the history of mankind, but the scientific notion of temperature is a theoretical construct, even though it has a close correspondence to subjective experience. The particular metrics chosen related to water boiling (in both Fahrenheit and Celsius), to Freezing (in Celsius), and to the "coldest" temperature that could be achieved with water, ice and salt (in Fahrenheit). Lord Kelvin also took the amazing step of developing a notion of temperature that is *really* abstract. His zero point of minus 273.15 degrees Celsius has never quite

been reached, and is far below what any person could experience. Yet it is very real as a scientific construct, one that is part of the fabric of physical science and ties various aspects of science together in that fabric.

Many other common terms in physical theory, like mass and gravity, are theoretical constructs, though they are related to human senses. Only in relatively recent physics history have mass and gravity been understood, and we owe that understanding to bits of inspiration on the part of Galileo and Newton. Having only half a century of AI history to look back on, we cannot really expect to have such a firm fabric of theoretical constructs stitched together. But some ideas are given below, after a comparison of Sciences that study natural and the artificial systems.

Sciences of the Artificial and their relation to Natural Sciences

Herbert Simon came to the conclusion that there was a place for what he called “Sciences of the Artificial” in his important book [69]. He did not *invent* the study of artifacts in a systematic manner, but he realized accurately and acutely that artifacts could be subjects of “real sciences”, with deep theories of the sort that exist in natural sciences. We will now consider some of the implications of this idea.

The boundaries between sciences of the artificial and the natural sciences are not clear-cut in practice because nature colors human artifacts, determining their possibility and their features. The “engineering sciences”, the portions of engineering that has been formalized in the sense of that they can predict the behavior of artifacts, including aspects such as stability and strength can be considered sciences of the artificial. The reason that this is not remarked upon more often is that they have called upon physical sciences more and more over the centuries to aid the “ingenuity” that gives the profession its name.

Linguistics is a science of the artificial. Human language is the artifact that it studies. But of course, the properties of the artifact are shaped by the natural properties of human learning and cognition, human hearing and speech in many ways. In the domain of phonetics, for example David Stampe’s “natural phonology” [Stampe 73, Donegan and Stampe 79] characterizes some of the interactions between language as an artifact and as a natural

phenomenon. We do not understand even yet the extent of the interaction between linguistics and human cognition. Is there an LAD (language acquisition device) [Chomsky 75] innate in humans that is specific to language, or is the learning of language based on the same principles as such other acquired systems as visual perception? Nobody knows for sure; but whatever the case, the nature of the world and the nature of learning processes must affect language.

Computer Science is a science of the artificial. Certainly, this is true insofar as it studies computers, which are artifacts; but also to the extent that it studies algorithms, which are human creations, too. The main subject studied in much of Computer Science is not computers but information, and the “state”, which is all the relevant information about a system at a given time, is therefore a fundamental theoretical construct. Information is a theoretical construct that is also fundamental in the natural sciences, but whose significance as a theoretical construct has only become apparent in this century, as its relationship to entropy and its role in quantum theory have been realized. So again, Computer Science has both artificial and natural parts.

Economics, another science of the artificial, studies a major artifact, the economy, and looking at this science of the artificial can provide some insight into the position of AI as a science of the artificial, and of the role of measurable theoretical constructs.

Predictive Measurement in a Science of the Artificial – An Example from Economics

Economics has struggled for longer than AI or computer science has existed to find theoretical constructs that have predictive power. It deals with large amounts of aggregated data, so the empirical data are statistical in nature. As of this date, economic theory is still not as crystal-clear as physics in terms of the role of its theoretical constructs, but its theoretical constructs, measured by expensively-gathered data by governments and multi-governmental agencies, are used regularly.

Recently, the U.S. Federal Reserve has been aggressive in raising interest rates because the *unemployment rate* (a construct measured by job creation and unemployment data) has been high and *economic growth* (a construct measured by GDP change and other data) has been rapid. In their models, these predict higher *inflation* (a

construct measured by PPI, CPI, and other data). Somewhere in the complex equations that describe the relationship among these theoretical constructs, and the construct inflation, it has recently been noticed that there is a need for the construct *productivity*. Economic theory must relate these constructs and others: average interest rates, demand for money and goods, money and commodity supply, savings rate, etc.

The definition of the constructs mentioned above is still hazy, and the relations among them are not mathematically precise. Some economic theories are incorporated in complex computer models. Their predictive value is not great, but they are getting better, and provide an example of the sort of prediction that is desirable for AI.

Surface Measures and Theoretical Constructs in AI – Some Examples

The sort of predictive ability that economists want, we would like to see in AI, too. If we have theoretical constructs at some deeper level, we can also use the theories of which they are a part to simulate or predict mathematically what happens if we increase or decrease parameters related to those constructs. It is a thesis of this paper that *there are theoretical constructs that can predict system performance measured in terms of surface measures*. At this point in the development of AI as science, it is hard to say just exactly what they would be, but some ideas can be drawn from today's AI and related subjects.

An Example Construct: Robustness

A surface measurement that could be very valuable across a variety of systems is some measure of *robustness* – the ability to exercise intelligent behavior over a large number of tasks and situations. From a computation-centered standpoint, if systems become robust, AI progress would be easier to see. From a cognition-centered standpoint, a system can never really be intelligent if it is not robust. (One way to think of a measure of intelligence in a single system would be as a measure of performance, robustness and autonomy.) The *surface* way to determine the robustness of a system would be to try it on a number of tasks and see how broad its methods are. But what *makes* intelligent systems robust? Learning ability, experience, and the ability to transfer that experience to new situations are all things that come to mind. A rough sketch of how measuring theoretical constructs in those areas

might give us a *predictive* figure for developing robust systems is given below.

Robustness: Learning?

If learning can make systems more robust, it should be interesting to measure the strength of the system's learning component. How easily does it adapt the system to a new situation? *Unsupervised learning* has wide applicability, but it can basically only determine clusters of similar items. *Supervised learning* must be presented with exemplars to learn relations, which seems not to be enough for a machine to extend its own capabilities. *Reinforcement learning* (RL) is a blend of both cognitive and computational centered AI. It started out as a model of classical conditioning, but turned out to be applied dynamic programming. There are a number of different techniques within RL, all of which have many possible applications. Neural nets or other approaches may be used. The theoretical constructs include the *state space* chosen, the *reinforcement function*, and the *policy*. The field is becoming quite sophisticated, and there are known facts about the relation of these to outcomes in particular cases [Mahadevan and Kaelbling 96]. Suppose that a reinforcement learning system constitutes a part of the intelligence of an intelligent system. There should be some way of predicting how that system would do upon encountering problems of a certain nature. By knowing how it chooses the concepts in its system and how they react on problems of that type, one can provide a partial evaluation of how effective the learning system would be. By obtaining such figures for all such subsystems, one could relate them to the performance of the full intelligent system. There is much work to be done in that direction.

Under certain circumstances, one can imagine learning extending robustness; but having to learn each new variations of a problem, even by reinforcement, is unlikely to lead to robustness quickly. It is expected that reinforced behaviors learned in one situation might be identical to those needed in another system, so this may lead to more rapid or better learning in the second situation. One approach to this is to condition behaviors that are not built into the system initially, as explored by Touretzky and Saksida [97]. But, still, one would like to have more general ways of reusing "big pieces" of learned knowledge.

Robustness: Transfer of Learning?

Transfer of learning is a phenomenon that we may be able abstract to theoretical constructs that can help to predict robustness. It is still not a deep measure, so it will then be important to predict transfer of learning from deeper constructs which will be mentioned below. At present, it is a research challenge to build transfer of learning into systems. But it is possible to see how one could test for it.

As far as measurement, here is roughly how transfer of learning might be measured:

1. Machine performance is measured on Task 1. The score is $P(t_1, T_1)$ = performance at time t_1 on Task 1. P is some suitably broad performance measure.
2. Performance is measured on Task 2 without learning (this being an artifact where we can control learning) to obtain $P(t_1, T_2)$ (keeping the time variable the same because the same machine abilities are assumed without learning even if the measurements are not simultaneous).
3. Note that if the measure is to have a meaning, previous training that might affect T_1 or T_2 must be controlled for, which could be difficult.
4. The machine is now allowed to perform task T_1 in which it learns, achieving better performance at some time t_2 , i.e. $P(t_2, T_1) > P(t_1, T_1)$.
5. It is then tested on T_2 , and the question is whether $P(t_2, T_2) > P(t_1, T_2)$ without having done additional learning on Task 2.

If indeed $P(t_2, T_2) > P(t_1, T_2)$ in some quantifiable way, the system has achieved (at least locally) one of the goals of AI, the transfer of learning from T_1 to T_2 . The amount of transfer can be measured by the amount of improvement on task2 as a function of the amount of training on task T_1 . Let us assume that we can describe this by some transfer effectiveness function, E for the system being tested. Let us say $E(T_1, T_2, t)$ gives “the effectiveness of training on T_1 for time t in terms of transfer to T_2 ”. We could describe this by a graph of performance on T_2 as a function of time being spent on T_1 .

Developing such a measure of transfer of learning and getting it accepted is not simple. To be useful, we would need a way of comparing T_1 and T_2 , to be sure that the second task is not just a subtask to the first. Difficult or not, defined measurements such as these are steps toward understanding the construct “transfer of learning”

and achieving it in artifacts. The measurable transfer construct would, in turn, help to provide a measurement of robustness, since learning transfer can make a system more robust. It is a step toward measurement of intelligence, at least by some definitions of intelligence, and, intuitively, at least, would have some predictive power.

How might we go about defining the similarity of T_1 and T_2 , as suggested above? We would have to decide what we mean by similarity of task. An interesting essay in this area is “Ontology of Tasks and Methods” [Chandrasekaran, Josephson and Benjamins [98]].

Various candidates for potentially measurable constructs that could be used to produce transfer but also to relate transfer to other phenomena are mentioned in a book edited by Thrun and Pratt [98], who have both had a research interest in learning-transfer processes. From the computation side comes the possibility of changing *inductive bias*. From the cognition-centered side, there is *generalization* from things already learned; but *overgeneralization* can be a major problem in learning, so it needs to be constrained. (Some simple constraints on overgeneralization in language learning are discussed in [Reeker 76].)

Robustness: Case-Based Reasoning?

Case-based reasoning is an intuitively appealing technique that was mentioned earlier in this paper. The idea is that one learns an expanding set of cases and stores the essentials of them away according to their conventional features. They are then retrieved when a similar case arises and mapped into the current case. Potential theoretical constructs include *indexing* and *retrieval* methods for the cases, case *evaluation* and case *adaptation* to the new situation. The cases could also be abstracted and generalized to various degrees, to a *model*.

Case-based reasoning is important for cognition centered AI. It is intuitively the way many people often figure out how to do things, and is thus embodied in the teaching methods of many professional fields – law, business, medicine, etc. It provides a launching pad for creativity as well, as mappings take place from one case to an entirely new one. Perhaps the new case is not really concrete, but a vague new idea. Then the mapping of an old case to it may result in a creative act – what we usually call

analogy. Analogy, *metaphor* in language, is a rich source – absolutely ubiquitous – of new meanings for words, and thus of new ways to describe concepts, objects, actions. Perhaps one key to robustness is the ability to use analogy. Four interesting papers by researcher in the area can be found in an issue of *American Psychologist* [Gentner *et al* 97].

Existing Surface and Subsurface Performance Measures

Researchers in text-based information retrieval (IR) have traditionally considered themselves not to be a part of the AI field, and some have even considered that artificial intelligence was a rival technology to theirs; but there is an overlap of interest. It is worth noting that IR has had a useful surface measure of system performance that has guided research and allowed comparison of technologies. The measure consists of two numbers, *recall* and *precision* [Salton 71]. Recall measures the completeness of the retrieval process (the percentage of the relevant documents retrieved). Precision measures the purity of the retrieval (the percentage of retrieved documents judged relevant by the people making the queries). If both numbers were 100%, all relevant documents in a collection would be retrieved and none of the irrelevant ones. Generally, techniques that increase one of the measures decrease the other. Real progress in the general case is achieved if one can be increased without decreasing the other.

For the IR community, better recall and precision numbers have both shown the progress of the field. They also show that it is still falling short, keeping up the challenge, especially as the need to use it for very large information corpora rises. In addition, they provide a standard within the community for judging various alternative schemes. Given a particular text corpus, one can consider various weighting schemes, use of a thesaurus, use of grammatical parsing that seeks to label the corpus as to parts of speech, etc., to improve the retrieval process. The interesting thing is to relate these methods and the characteristics of the corpus to precision and recall, but so far that has not been sharp enough to quantify generally.

Related to information retrieval is automated natural language information extraction, which tries to find specified types of information in bodies of text (often to create formatted databases where extracted information can be

retrieved or mined more readily). A related but different (cost-based) measure was defined several years ago for a successful information extraction project [Reeker, Zamora and Blower 83]. One measure was *robustness* (over the texts, not different tasks as in the broader intelligent systems usage discussed earlier). This was defined as the percentage of documents out of a large collection that could be handled automatically. The idea was that some documents would be eliminated through automated pre-screening (because those documents were not described by the discourse model the system used) and relegated to human processing. Another measure was *accuracy* (the percentage of documents not eliminated that were then correctly processed in their entirety, by the system). Yet another was *error rate* (the percentage of information items that were erroneous – including omitted - in incorrectly handled documents). From this more detailed breakdown, estimates of the basic cost of processing the documents, based on human and machine processing costs and costs assigned to errors and omissions, was derived. The measure could be used to drive improvements in information extraction systems or decide whether to use them, compared to human extraction (which also has errors) or to improve the discourse model to handle a larger portion.

For information extraction projects, it was further suggested that the cost of erroneous inputs might drive a built-in “safety factor” that could be varied for a given application [Reeker 85]. This safety factor was based on linguistic measures of the text (in addition to the discourse model) that could cause problems for the system being studied. The adjustable safety factor could be built into the prescreening mentioned above. In other words, the system would process autonomously to a greater or lesser degree and could invite human interaction in applications where the cost of errors was especially high. It was suggested that the system would place “warning flags” to help it make a decision on screening out the document, and these could also aid the human involved. Although this was a tentative piece of work, the idea of tying a surface measure (robustness) into the underlying properties of the system is exactly like tying measurable surface properties into underlying theoretical constructs. The theoretical constructs mentioned in this case were structural or semantic ones from linguistics.

From the area of software engineering comes another tradeoff measure that is worth mention. The author did some work on ways of providing metrics - surface metrics, initially - for program readability (or understandability) [Reeker, 79]. Briefly, studies of program understanding had identified both go-to statements and large numbers of identifiers (including program labels) as problems. At the same time, the more localized loop statements could result in deep embeddings that were also difficult to understand for software repair or modification. The vague concept of readability could be replaced by a measure of go-to statements and maybe also one of the number of different identifiers. This particular study suggested *depth of embedding* as a problem and also suggested a tradeoff between depth of embedding a metric called *identifier load*. Identifier load was a function of the number of identifiers and the span of program statements over which they were used. Identifier load tended to increase as depth of embedding was reduced by the obvious methods.

There were a number of similar software metrics studies in the 1970s, and they continue. This approach, however, was part of an attempt to look at natural language for constructs that might be of relevance in programming languages and programming practice [Reeker 80]. The *depth* measure was based on an idea of Victor Yngve [60], which came out of his work in linguistics - an idea that retains a germ of intuitive truth. Yngve had in turn related his natural language measure of embedding depth to measures of short-term memory from cognitive psychology. Whether these relationships turn out to be true or lead to related ideas that are true or not, they illustrate how theoretical constructs can stitch AI, computer science, and other artificial and natural sciences together. They also illustrate the quest for metrics that can firm up the foundations of the sciences.

More Constructs To Be Explored

There are many more existing theoretical constructs that have arisen within AI or been imported from computer science or cognitive science that beg to be better defined, quantified, and related to other constructs, both deep and surface.

Means-ends analysis and *case based reasoning* have both been mentioned as forms of problem solving. How do these cognitive characterizations of problem solving relate to one another? At a deeper level is the construct

of *short term memory* mentioned in the previous section in relationship to Yngve's *depth*. How does short-term or working memory relate to long term memory and how are the two used in problem solving? The details are not known. The size of a short-term memory may not be as relevant in a machine, where memory is cheap and fast. But we cannot be sure that it is not relevant to various aspects of machine performance because it is reflected at least in the human artifacts that the machine may encounter. For instance, in resolving anaphora in natural language the problem may be complicated if possible referents are retrieved from arbitrarily long distances.

A similar problem arises from long-term memory if everything ever learned about a concept is retrieved each time the concept is searched for. This can lower retrieval precision (to use the term discussed earlier for machine retrieval) and cause processing difficulties on a given problem. It may be that Simon's notion of *bounded rationality* is a virtue in employing intelligence. Are we losing an important parameter in intelligence if we try always to optimize rationality? For AI system, *anytime algorithms* and similar constructs for approximate, uncertain, and resource bounded reasoning have been developed in recent years, and hold a good deal of promise [Zilberstein 96].

An interesting theoretical construct arising out of AI knowledge representation and the attempts to use it in expert systems and agents and for other purposes is that of an *ontology*. "Ontology" is an old word in philosophy designating an area of study. In AI it has come to designate a type of artifact in an intelligent system: The way that that system characterizes knowledge. In humans, ontologies are shared to a large degree, but certainly differ from every person to every other, despite the fact that we can understand each other. Are some ontologies indicative of more intelligence than others in ways that we can measure? One suggested criterion for high intelligence is the ability to understand and use very fine distinctions (or to actually create new ones, as described in Godel's memorandum cited by Chandrasekaran and Reeker [74]). Is an ontology's size important, or its organization, or both? Can one quantify a system's ability to add new distinctions?

A related issue is *vocabulary*. Many people think that an extensive vocabulary, *used appropriately*, is a sign of intelligence, or at least scholastic aptitude. In computer programs that

do human language processing, the vocabulary consists of a *lexicon* that generally also has structural (*syntactic*) information for parsing or generating utterances containing the lexical item and *meaning representations* for the lexical item. The lexicon can be much larger than any human's vocabulary; but for the vocabulary to be used appropriately for language production or understanding, it still falls far short of the human vocabulary. For that to be improved better techniques of *semantic mapping* are required, including links to ontologies and methods of inferring the ontological connections and of idiosyncratic aspects of speakers with which a conversation is taking place. Is the vocabulary an indication of the size of the ontology and the distinctions it makes, or vice-versa? Nobody knows; but better theories of how they link up are needed for both understanding and fully effective use of human language by intelligent systems.

Another cognitive concept that is still a mystery is *creativity*, certainly a part of intelligence, or at least of high intelligence. Does the ability to add entirely new concepts, not taught, constitute creativity? How does one harness serendipity to develop creativity? Is creativity linked with *sensory cognition*, the cognitive phenomena related to senses, such as vision, including perception, visual reasoning, etc. There is a need for deep theoretical constructs underlying notions like creativity, and for measures of these constructs and their attributes [Simon 95, Buchanan 00].

Turning to computational constructs, we notice that much of the AI described above takes place through various forms of *search*. Already there exists a pretty good catalogue of variations on search and how to manage it, in which a good deal of theory is latent. Some of the search is of a *state space*, involving the ubiquitous state concept basic to theoretical computer science. Search is also coupled with *pattern matching*, which underlies many of the methods mentioned earlier in this paper.

The potential constructs mentioned here are just a sample of the ones already available in Artificial Intelligence, and to them should be added others found in some of the major works of Newell and Simon on Problem Solving and Cognition [Newell and Simon [65], Newell [87]].

Summary and Author's Note

The development of a true science of artificial intelligence is something that has concerned the author for a long time. It has been encouraging to see the development within the field of interesting and non-obvious theoretical constructs. This paper has suggested that theoretical constructs with attributes that we can measure are especially valuable and it has suggested a number of such candidates. The paper suggests that we enlist Lord Kelvin's emphasis on measurement in choosing such constructs. These same measurable theoretical constructs will in many cases relate (at least at deeper levels) to those of cognitive science, computer science, and other sciences. They will help predict measures at the surface that can be used to provide metrics for the performance (and through that, the intelligence) of intelligent artifacts. We should have in mind the quest for such measurable constructs as we move forward in creating intelligent artifacts.

References

- Buchanan, B. G. [00], Creativity at the Meta-Level, Presidential Address, American Association for Artificial Intelligence, August 2000. [Forthcoming in *AI Magazine*.]
- Chandrasekaran, B., J. R. Josephson and V. R. Benjamins [98] Ontology of Tasks and Methods, 1998 Banff Knowledge Acquisition Workshop. [Revised Version appears as two papers "What are ontologies and why do we need them?," *IEEE Intelligent Systems*, Jan/Feb 1999, 14(1); pp. 20-26; "Ontology of Task and Methods," *IEEE Intelligent Systems*, May/June, 1999.]
- Chandrasekaran, B. and L. H. Reeker [74]. "Artificial Intelligence – A Case for Agnosticism," *IEEE Trans. Systems, Man and Cybernetics*, January 1974, Vol. SMC-4, pp. 88-94.
- Chomsky, Noam [65]. *Aspects of the Theory of Syntax*. MIT Press, Cambridge MA.
- Chomsky, N. [75] *Reflections on Language*. Random House, New York.
- Donegan, P. J. & D. Stampe [79]. The study of Natural Phonology. In Dinnsen, Daniel A. (ed.). *Current Approaches to Phonological Theory*. Indiana University Press, Bloomington, 126-173.
- Ernst, G. & Newell, A. [69]. *GPS: A Case Study in Generality and Problem Solving*. Academic Press, New York.

- Fitch, F. B. [52]. *Symbolic Logic*. Roland Press, New York.
- Gentner, D., K. Holyoak *et al* [97]. Reasoning and learning by analogy. (A section containing this introduction and other papers by these authors and A. B. Markman, P. Thagard, and J. Kolodner, *American Psychologist*, 52(1), 32-66.)
- Gentzen, G. [34] Investigations into logical deduction. *The Collected Papers of Gerhard Gentzen*, M. E. Szabo, ed. North-Holland, Amsterdam, 1969. [Published in German in 1934.]
- Kolodner, J.L. [88] Extending Problem Solving Capabilities Through Case-Based Inference, In, *Proceedings of the DARPA Case-Based Reasoning Workshop*, Kolodner, J.L. (Ed.), Morgan Kaufmann, Menlo Park, CA.
- Leake, D. B. Ed. [96] *Case-Based Reasoning: Experiences, Lessons, And Future Directions* Indiana University, Editor 1996, AAAI Press/MIT Press, Cambridge, MA.
- Margenau, H. [50]. *The Nature of Physical Reality*, McGraw-Hill, New York.
- Mahadevan, S. and L. P. Kaelbling [96] The National Science Foundation Workshop on Reinforcement Learning. *AI Magazine* 17(4): 89-93.
- Meystel, A. *et al* [00] Measuring Performance of Systems with Autonomy: Metrics for Intelligence of Constructed Systems. *In this volume*.
- Newell, A. [87] *Unified Theories of Cognition*. Harvard University Press. Cambridge, Massachusetts, 1990. [Materials from William James Lectures delivered at Harvard in 1987.]
- Newell, A., and H.A. Simon [63] GPS, a program that simulates human thought, *Computers and Thought*, E.A. Feigenbaum and J. Feldman (Eds.), McGraw-Hill, New York.
- Newell, A., & Simon, H.A. [65]. Programs as theories of higher mental processes. R.W. Stacy and B. Waxman (Eds.), *Computers in biomedical research* (Vol. II, Chap. 6). Academic Press, New York.
- Newell A. & Simon H.A. [72] *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, NJ.
- Reeker, L. The computational study of language acquisition, *Advances in Computers*, 15 (M. Yovits, ed.), Academic Press, 181-237, 1976.
- Reeker, L. [79] Natural Language Devices for Programming Language Readability; Embedding and Identifier Load, *Proceedings, Australian Computer Science Conference*, Hobart Tasmania.
- Reeker, L., E. Zamora and P. Blower [83] Specialized Information Extraction: Automatic Chemical Reaction Coding from English Descriptions, *Proceedings of the Symposium on Applied Natural Language Processing*, Santa Monica, CA, Association for Computational Linguistics, 1983.
- Reeker, L. H. [80] Natural Language Programming and Natural Programming Languages, *Australian Computer Journal* 12(3): 89-93.
- Reeker, L. H. [85] Specialized information extraction from natural language texts: The "Safety Factor", *Proceedings of the 1985 Conference on Intelligent Systems and Machines*, 318-323, Oakland University, Michigan, 1985.
- Salton, G. [71] *The SMART Retrieval System*, Prentice-Hall, Englewood Cliffs, NJ.
- Simon, H. A. [69] *The Sciences of the Artificial*. Third Edition. Cambridge, MA, MIT Press, 1996. [Original version published 1969].
- Simon, H. A. [95] Explaining the Ineffable: AI on the Topics of Intuition, Insight and Inspiration. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95*, Montréal, Morgan Kaufmann, Menlo Park, CA, Volume 1, 939-949.
- Stampe, D. [73] *A Dissertation on Natural Phonology*. New York: Garland Publishing, 1979. [Original University of Chicago dissertation submitted in 1973.]
- Thrun, S. and L. Pratt (eds.) [98]. *Learning To Learn*. Kluwer Academic Publishers.
- D.S. Touretzky and L.M. Saksida [97] Operant conditioning in Skinnerbots. *Adaptive Behavior* 5(3/4):219-247.
- Turing, A. M. [50]. "Computing Machinery and Intelligence." *Mind* LIX (236 ; Oct. 1950): 433-460 reprint in [*Collected Works of A. M. Turing* vol. 3: Mechanical Intelligence, D. C. Ince ed., Elsevier Science Publishers, Amsterdam, 1992: 133-160].
- Weizenbaum J [66]. ELIZA - a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9, 36-45.
- Weizenbaum, J. [74]. Automating psychotherapy. *Communications of the ACM*, 17(7):425, July 1974.
- Yngve, V [60] The depth hypothesis, *Proceedings, Symposia in Applied Mathematics*, Vol. 12: Providence, RI, American Math. Society, 1961. [Based on publication under another title, 1960.]
- Zilberstein, S. [96] Using Anytime Algorithms in Intelligent Systems, *AI Magazine*, 17(3):73-83, 1996.

Intelligence with Attitude

W. C. Stirling and R. L. Frost

Electrical and Computer Engineering Department
Brigham Young University
Provo, UT 84602

ABSTRACT

An essential feature of intelligence is the ability to make autonomous choices. A new paradigm of satisficing decision making incorporates two utilities for decision making, rather than the usual single utility that is characteristic of optimal decision making. These two utilities may be used to define figures of merit for the intellectual power or fitness of the decision maker as it functions in its environment. These utilities may also be applied in group settings. In particular, societies of negotiatory decision makers may undergo considerable tension as they attempt to reach a compromise that is acceptable to the group as a whole and to all members of the group.

KEYWORDS: *multi-agent decision theory, satisficing, attitude, negotiation*

1. INTRODUCTION

There are three issues that must be addressed in the design of an intelligent decision system: (a) defining the alternatives, (b) defining the preferences, and (c) choosing between the alternatives as a function of the preferences. The first two issues are highly dynamic. Alternatives may appear and disappear and preferences may change. Much of the study of intelligent systems is properly focused on these dynamics. At the moment of truth when a decision must be made, however, we must assume that the alternatives and preferences have been defined, and all that remains is to make the choice. This paper focuses on this last, consummate step.

The ability to make decisions is essential to intelligent behavior. Indeed, the word *intelligent* comes from the Latin roots *inter* (between) + *legere* (to choose). We thus assume that there is only one essential characteristic of intelligence in man or machine—an ability to choose between alternatives.

Choices between alternatives, or decisions, are usually justified by the maximization of expected utility, an approach Simon calls *substantive rationality* [8]. We argue that for multiple agents, especially those in dynamic environments, the requirement for substantive rationality is too demanding. First, although a solution may exist, the information or computing power necessary to find it may be unavailable. We will often be

forced to fall back on what Simon terms *procedural rationality*, or the reliance on heuristic or *ad hoc* procedures defined by an authority. Second, and more serious, is that the existence of an optimal solution may be in doubt. Von Neumann-Morgenstern game theory shows that for many games a solution that is simultaneously best for the group and for each individual in the group simply does not exist. This seems to imply that a theory of group decisions satisfactory for the synthesis of coordinating agents cannot be obtained by a straightforward maximization of utility.

We are thus motivated to consider definitions of rationality upon which we can build a more robust theory of intelligent multi-agent decision making. We hold that the fundamental obligation of a rational decision maker is to make decisions that are, in some well-defined sense, good enough. Historically, the study of good enough decisions was first formalized by Simon, when he introduced the term *satisficing* to characterize decisions that achieve the decision maker's *aspiration level* [6, 7]. This notion of satisficing defines quality according to the criteria used for substantive rationality, but evaluates quality against a standard that is chosen more or less arbitrarily. It essentially blends substantive and procedural rationality, and is a species of what is often termed *bounded rationality*.

Rather than blend the two extremes of substantive and procedural rationality *a la* Simon, our work explores an alternative which leads naturally to a set of satisficing solutions that is consistent with Simon's intent. It also guarantees the existence of jointly rational decisions, and seems to be a natural vehicle for the design and synthesis of intelligent decision systems.

We start by assuming that the most primitive way to make decisions is to make intra-option comparisons in the form of dichotomies. We define two distinct (and perhaps conflicting) sets of attributes for each option and to either select or reject the option on the basis of comparing these attributes. Such dichotomous comparisons are *intrinsic*, since the evaluation of an option's merits is not referenced to anything not directly related to the option, including other options. They are also local comparisons; it is not possible to form a global ordering the options on the basis of such comparisons. An *intrinsically rational* choice is one for which the decision maker's benefits are at least as great as its costs. We define a *satisficing decision*

as one that is intrinsically rational,¹ because these options are good enough, in the sense that their attributes have been favorably compared with a standard. We differ from Simon only in the standard used for comparison: the positive and negative attributes of each option, versus externally supplied aspiration levels.

Intrinsic rationality appears to be a weaker notion than substantive rationality. Although it identifies all options that are, in the sense we have defined, good enough, it does not insist on a unique solution. At the moment of truth, the decision maker may choose any of the satisficing options with the assurance that it will at least get its “money’s worth.” In practice, however, the advantage of a theory founded on substantive rationality may be more illusory than real. Objective functions themselves are often created by an *ad hoc* combination of preferences into a single performance index, and this combination can be, and usually is, manipulated until satisfactory behavior is achieved. Thus, even optimization approaches rely in their application on satisficing notions, however informally.

As mentioned earlier, our approach to intrinsic rationality requires the definition of two preference functions, one to characterize the desirable attributes, and one the undesirable attributes, of each option. An option is desirable to the degree that it achieves the goal. It is undesirable to the degree to which its adoption consumes the decision maker’s resources, such as energy, safety, or other costs. Separate preference functions permit the development of metrics to evaluate how suited the decision maker is to function in its environment. Intuitively, if a decision maker has options available to it that achieve its goal with low cost, it is well-suited for its environment. On the other hand, if it must incur great cost or undergo great risk to achieve its goal, it is clearly not as well suited. Although the goal may be achieved equally well in either case, there is a fundamental difference in the ability of the agent under the two scenarios. This difference may not be easily discernible under the substantive or procedural rationality paradigms, but it is clearly discernible under the intrinsic rationality paradigm.

In the following we first summarize the mathematical development of satisficing decision theory. We next introduce a concept of *attitude*, or disposition, for the agents, and develop figures of merit for evaluating the equivocation experienced by the decision maker or decision making system. We then present a basic negotiation theorem and describe a simple negotiatory process to converge to a rational compromise. We then finish with an example and draw conclusions.

2. SATISFICING

Von Neumann-Morgenstern game theory is based on a very sophisticated paradigm—global optimization. There are a number of basic problems, however, with optimization-based ap-

¹Other researchers have appropriated this term to describe various notions of constrained optimization. In this paper, we restrict our usage to be consistent with Simon’s original concept.

proaches. First, since it is well known that humans are not good optimizers [1, 2, 5], a decision-making system that seeks to approximate human behavior may be unnecessarily constrained by insisting on, and only on, optimal performance. Second, optimization is a fixed, or absolute concept, in the sense that if an option is not the best, then it is unacceptable. There cannot be degrees of optimization. Third, optimization is, fundamentally, a notion of exclusive self interest, and does not easily generalize to settings where it is important to accommodate both group and individual interests [4]. It is usually impossible to arrive at a joint solution that is simultaneously best for the group as a whole and for each member of the group.

Our notion of satisficing, on the other hand, does not insist upon optimal performance, and in return for this concession it logically permits degrees of satisficing and the accommodation of both group and individual interests. By adjusting the tradeoff standards between cost and benefit, it may be possible to find a joint solution that is simultaneously good enough for the group and good enough for each member of the group. This is the fundamental goal of negotiation.

Our approach is to employ the mathematics, but not the usual semantics, of probability theory. As discussed in [9, 10] we may encode the preference relationships via mass functions, which we term the *selectability* and *rejectability* functions. By so doing, we are able to account for conditional preferences (analogous to conditional probabilities) and to express both joint (group) and marginal (individual) preferences.

We formalize this procedure as follows. Let U_i denote the option set for the i th agent (we will assume U_i is of finite cardinality), $i = 1, \dots, N$, let $\mathbf{U} = U_1 \times \dots \times U_N$ denote the product space of joint options, and let $\mathbf{u} = \{u_1, \dots, u_N\}$, where $u_i \in U_i$, denote an option vector. Let $p_S(\mathbf{u})$ indicate the degree to which the joint option \mathbf{u} is successful in achieving a group goal. We require that $\sum_{\mathbf{u} \in \mathbf{U}} p_S(\mathbf{u}) = 1$ and $p_S(\mathbf{u}) \geq 0$, so p_S is a mass function, which we term the *joint selectability mass function*. Also, let $p_R(\mathbf{u})$ indicate the degree to which the joint option \mathbf{u} consumes resources, and require this to also be a mass function, which we will term the *joint rejectability mass function*. Next, let $p_{S_i}: U_i \rightarrow [0, 1]$ and $p_{R_i}: U_i \rightarrow [0, 1]$ be marginal selectability and rejectability mass functions, respectively, derived from p_S and p_R by appropriate summation. For a discussion of how these joint and marginal mass functions may be practically constructed, see [9, 10].

These mass functions define a dichotomy for each option, that is, they partition the attributes of the option into two categories and provide a measure of support for each class of attributes. We evaluate each dichotomy by comparing the selectability (benefit) to the rejectability (cost) of each option. By so doing, we define the *jointly satisficing set*

$$\Sigma_b = \{\mathbf{u} \in \mathbf{U}: p_S(\mathbf{u}) \geq b p_R(\mathbf{u})\},$$

and define the *individually satisficing sets*

$$\Sigma_b^i = \{u \in U_i: p_{S_i}(u) \geq b p_{R_i}(u)\},$$

$i = 1, \dots, N$. The *boldness parameter*, b , is a constant in the interval $[0, 1]$, which is nominally set to unity, but may be decreased under special circumstances to be discussed below. Σ_b is the set of all joint options that are good enough for the group, and each Σ_b^i is the set of all individual options that are good enough for the i th agent.

These sets provide the agent or group of agents with the ability to make individual or group decisions. If the i th individual agent is empowered to make its own decision, it may choose any member of Σ_b^i . If the group as a whole is to make a collective decision, it may choose any member of Σ_b . These choices may be random, or they may be made according to some tie-breaking procedure.

3. EQUIVOCATION

Human decision makers often make qualitative assessments of the difficulty, in terms of stress or tension, encountered in making decisions. Even if such knowledge does not have a direct bearing on their immediate decisions, an appreciation of the difficulty involved in forming the decision is an important aspect of the decision-making experience. A decision maker need not possess anthropomorphic qualities, however, to assess the difficulty of making decisions, and we do not propose to endow an artificial decision maker with some sort of ersatz anthropomorphic capability. Under our satisficing approach, however, it is possible to evaluate attributes of the decision problem that correspond more to its functionality and fitness than to its success.

Are decisions easily made and implemented, or do they tax the capabilities of the decision maker? Such assessments are not a typical undertaking of classical decision theory. Maximizing expectations has no need to concern itself with issues such as “difficulty.” Nevertheless, choices are not all of equal difficulty.

By employing two utilities, rather than only one, we may analyze them to ascertain the compatibility of the attributes of the preferences. If they are compatible, in that options that conserve resources also achieve the goal, then the decision maker is in a fortunate situation of being content. If the preferences are incompatible, in that options that achieve the goal also are highly consuming of resources, then the decision maker is fundamentally conflicted. These attributes constitute attitudes, or dispositions, of the decision maker.

The optimization literature is devoid of discussions concerning the attitude or disposition of the decision maker who, like the paradigm it employs, is assumed to be dispassionate. It is simply doing what should be done under the auspices of individual rationality, and attitudes or feelings, should they even exist (and they need not), are completely irrelevant. Furthermore, to attribute anthropomorphic characteristics to a decision maker would be seen by many as nothing more than a concocted story line that is of marginal value if not completely misleading.

3.1. Attitude

It is fortunate if an option that conserves resources (low rejectability) also achieves the goal (high selectability)—in this environment, a decision maker is content. Many interesting decision problems, however, are such that actions taken in the interest of achieving the goal are expensive, hazardous, or have other undesirable side effects. A decision maker in this situation is conflicted. Contentment and conflict are basic dispositional states that serve as guides to the decision maker’s functionality. A situation requiring frequent high-conflict decisions indicates that the tasks are difficult for the decision maker. Making high-conflict decisions, however, is not a measure of how well the decision maker is performing—it may, in fact, be making good, but costly, decisions. It is also true, however, that a high-conflict environment may result in poor performance because the decision maker is simply not powerful enough to deal adequately with its environment. Such a situation might serve as a trigger to prompt changes, such as activating additional sensors, or otherwise seeking more information about the environment. It may also trigger a learning mechanism to prompt the decision maker to adapt itself better to the environment.

Since selectability and rejectability are probabilities, it may be useful to appropriate some of the mathematical machinery of probability theory to aid in interpreting these quantities. One way to gain some insight is to examine the entropy of selectability and rejectability.

Definition 1 The **entropy** of a mass function p is

$$H(p) = - \sum_{u \in U} p(u) \log_2 p(u).$$

□

Entropy is usually employed in Shannon information theory as a measure of how much uncertainty (randomness or disorder) is reduced, on average, as a result of conducting an experiment governed by the mass function [3]. In our context, however, we wish to provide entropic interpretations for selectability and rejectability that are distinct from the usual probabilistic interpretation.

In assessing selectability, we consider expediency as analogous to uncertainty. To motivate this interpretation, suppose u' is implemented. If $p_S(u') \approx 1$, then $\log_2 p_S(u') \approx 0$ which is consistent with the notion that little reduction in expediency occurs if an option with high selectability is implemented. Conversely, suppose $p_S(u') \approx 0$, but is nevertheless implemented. Then $-\log_2 p_S(u')$ is large, indicating a great loss in expediency. The entropy of selectability is the average reduction in expediency that obtains as result of making choices according to p_S .

To interpret the entropy of p_R , we consider expense as analogous to uncertainty. Suppose u' is implemented. If $p_R(u') \approx 1$, then $\log_2 p_R(u') \approx 0$ which is consistent with the notion that little reduction in expense occurs if a highly rejectable option is nevertheless implemented. On the other hand,

if $p_R(u') \approx 0$ and u' is implemented, then $-\log_2 p_R(u')$ is large, indicating a great reduction in expense. The entropy of rejectability is the average reduction in expense that obtains as a result of making choices according to p_R .

Entropy is maximized by the uniform distribution; that is, if $p^*(u) = \frac{1}{n}$ for all $u \in U$, then $H(p^*) \geq H(p)$ for all mass functions p over U , and has entropy $H(p^*) = \log_2 n$. A uniform p_S generates the highest possible average expediency, and a uniform p_R would generate the highest possible average expense. Consequently, it is useful to take the uniform distribution as a baseline against which to assess the properties of arbitrary mass functions. Let n be the cardinality of the action space, U (assumed to be finite for this discussion).

Definition 2 If $p_S(u) = \frac{1}{n}$ (that is, selectability under p_S is equal to selectability under the uniform distribution), then the option is **success neutral**. If the selectability mass function is uniform, then the decision maker's attitude will be success neutral. \square

Definition 3 If $p_R(u) = \frac{1}{n}$ (that is, rejectability under p_R is equal to rejectability under the uniform distribution), then the option is **conservation neutral**. If the rejectability mass function is uniform, then the decision maker's attitude will be conservation neutral. \square

Definition 4 If $p_S(u) > \frac{1}{n}$ (that is, selectability under p_S is greater than selectability under the uniform distribution), then the option is attractive with respect to performance relative to other options— u is **expedient**. \square

Definition 5 If $p_R(u) > \frac{1}{n}$ (that is, rejectability under p_R is greater than rejectability under the uniform distribution), then u is unattractive with respect to cost or other penalty— u is **expensive**. \square

The relationship between selectability and rejectability permits the definition of four dispositional modes of the decision maker with respect to each of its options. Let U be the set of all possible options.

Definition 6 If $u \in U$ is both expedient and expensive, then the decision maker will desire to reject, on the basis of cost, an option that is suitable in terms of performance—it will be **ambivalent** with respect to u . \square

Definition 7 If $u \in U$ is both inexpedient ($p_S(u) < \frac{1}{n}$) and inexpensive ($p_R(u) < \frac{1}{n}$), then the decision maker will be desirous of accepting the option on the basis of cost, but will be reluctant to do so because of poor performance. The decision maker will be **dubious** with respect to u . \square

Definition 8 If $u \in U$ is expedient and inexpensive, then the decision maker is in the position of desiring to implement

an option that would yield good performance—a dispositional mode of **gratification** with respect to u . \square

Definition 9 If $u \in U$ is inexpedient and expensive, then the decision maker will desire to reject, on the basis of cost, an option that also provides poor performance, and will thus be in a dispositional mode of **relief** with respect to u . \square

These four modes provide a qualitative measure of the way the decision maker is matched to its task. Gratification and relief are modes of contentment, while dubiety and ambivalence are modes of conflict. Figure 1 illustrates these regions.

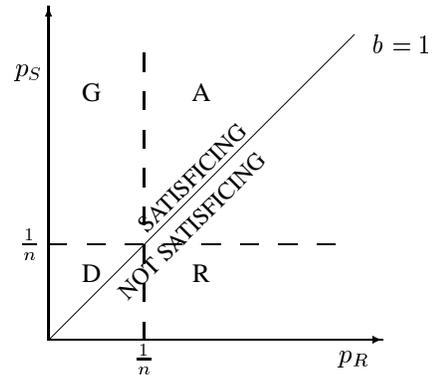


Figure 1: Dispositional regions: G = gratification, A = ambivalence, D = dubiety, R = relief.

Figure 2 illustrates various cases for $n = 2$, a two-dimensional decision problem. In these plots, the diagonal line represents the unit simplex, and the p_S and p_R values are plotted as vectors that lie on the simplex.

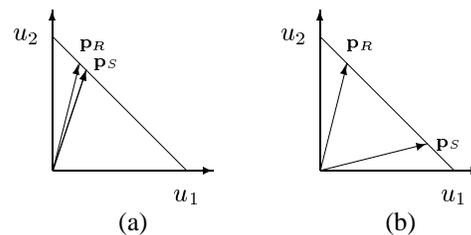


Figure 2: Attitude: (a) The decision maker is dubious with respect to u_1 and ambivalent with respect to u_2 . (b) The decision maker is gratified with respect to u_1 and relieved with respect to u_2 .

3.2. Figures of Merit

It would be useful to obtain formal expressions to capture some of the features of the qualitative analysis described in Section

3.1., where it is qualitatively indicated that as these distributions become more closely aligned, the decision maker becomes more ambivalent and dubious. We propose two measures that are similar, but not identical.

Diversity One important feature of the selectability and rejectability functions, therefore, is their dissimilarity. To obtain such a measure, we again appeal to the notion of entropy, and apply the Kulback-Leibler distance measure.

Definition 10 The **Kulback-Leibler (KL) distance measure** of two mass functions, say p_1 and p_2 , is given by

$$D(p_1 \parallel p_2) = \sum_{u \in U} p_1(u) \log_2 \frac{p_1(u)}{p_2(u)}.$$

□

The KL distance measure is an indication of the relative entropy of two mass functions. $D(\cdot \parallel \cdot)$ is not a true metric; it is not symmetric and does not obey the triangle inequality. It is, however, non-negative, and it is easily seen that $D(p_1 \parallel p_2) = 0$ if and only if $p_1(u) = p_2(u)$ for all $u \in U$.

We may apply the KL distance measure to the problem of ascertaining dissimilarity of the selectability and rejectability functions by computing the KL distance between selectability and rejectability.

Definition 11 The **diversity functional** is:

$$D(p_S \parallel p_R) = \sum_{u \in U} p_S(u) \log_2 \frac{p_S(u)}{p_R(u)},$$

or, equivalently,

$$D(p_S \parallel p_R) = - \sum_{u \in U} p_S(u) \log_2 p_R(u) - H(p_S).$$

□

Small values occur when the selectability and rejectability functions are similar, indicating a condition of potential conflict. If they are identical, then the decision maker is in a position of wishing to reject precisely the options that are in its best interest—an unfortunate condition of total paralysis.

Diversity is infinite if there exist options with nonzero selectability and zero rejectability. Such options are free options, since no cost independent of achieving the goal is incurred by adopting them (analogy: coasting saves fuel, but may or may not get you to your destination). Diversity is not a measure of performance; that is, if one decision maker has a more diverse selectability/rejectability pair than another, that is not an indication that it will perform better than the other. It does, however, provide an assessment of the environment in which the decision maker operates.

Tension Although the diversity functional provides insight into the relationship between selectability and rejectability, it does not afford a convenient comparison in the case where the decision maker is neutral with respect to either selectability or rejectability. To develop such a measure, it is convenient to re-normalize the selectability and rejectability functions. Consider first the case where p_S and p_R are mass functions and U is finite. Let

$$\begin{aligned} \mathbf{p}_S &= [p_S(u_1), \dots, p_S(u_n)] \\ \mathbf{p}_R &= [p_R(u_1), \dots, p_R(u_n)] \end{aligned}$$

be selectability and rejectability vectors, and let $\mu = [\frac{1}{n}, \dots, \frac{1}{n}]$ denote the uniform mass function vector, where n is the cardinality of U . Although these vectors are unit-length under the L_1 norm, they are not of unit length under the L_2 norm. It will be convenient to normalize these vectors with respect to L_2 . Let $|\mathbf{p}_S| = \sqrt{\mathbf{p}_S \mathbf{p}_S^T}$, with similar definitions for $|\mathbf{p}_R|$ and $|\mu|$. The L_2 normalized mass function vectors will be denoted by $\tilde{\mathbf{p}}_S = \frac{\mathbf{p}_S}{|\mathbf{p}_S|}$, and similarly for $\tilde{\mathbf{p}}_R$ and $\tilde{\mu}$.

We express the similarity between p_S and p_R through the inner product of the corresponding unit vectors, yielding the expression $\tilde{\mathbf{p}}_S \tilde{\mathbf{p}}_R^T$. This quantity will be unity when $p_S \equiv p_R$, and will decrease as the two mass functions tend toward becoming orthogonal, and thus captures some of the properties we desire to model. If we normalize by the product of the projections of \mathbf{p}_S and \mathbf{p}_R onto the uniform distribution, we tend to scale up the inner product as the mass function vectors become distanced from the uniform distribution.

Definition 12 The **tension functional** is

$$T(p_S \parallel p_R) = \frac{\tilde{\mathbf{p}}_S \tilde{\mathbf{p}}_R^T}{\tilde{\mathbf{p}}_S \tilde{\mu}^T \tilde{\mathbf{p}}_R \tilde{\mu}^T},$$

which simplifies into the convenient form:

$$T(p_S \parallel p_R) = n \mathbf{p}_S \mathbf{p}_R^T = n \sum_{i=1}^n p_S(u_i) p_R(u_i).$$

□

Clearly, $T(p_S \parallel p_R)$ is positive and bounded by the dimension, n . If either the selectability or rejectability is uniform, then the tension function equals unity. If the rejectability function is uniform, then the decision maker is rejectability-neutral. If the selectability is uniform, then the decision maker is selectability-neutral. If $T(p_S \parallel p_R) > 1$, then the projection of selectability onto rejectability is significant, and options that are desirable are also costly. We may interpret this as a state of conflict. On the other hand, if $T(p_S \parallel p_R) < 1$, then the projection of selectability onto rejectability is small, and the decision maker is in a state of contentment.

A decision maker operating in a contented environment is well-tuned to its task—decisions that possess high rejectability also possess low selectability. Such a decision maker should be

expected to achieve its goals with ease, and be adequate in most situations. A conservation-neutral decision maker will function much as would a conventional Bayesian decision-maker. If it is success-neutral, it will function much like a minimax decision-maker. If the decision maker is both conservation-neutral and success-neutral, it is completely indifferent to the outcome, and there is little point in even attempting to make a decision other than a purely random guess.

4. NEGOTIATION

Negotiation under the individual rationality paradigm forbids any individual participant, as well as any potential coalition, from settling for a decision that is below its security, or minimax, level. This is a very strong restriction, which can lead to an empty core and the lack of a rational basis for negotiation. There are many ways to modify this solution concept to justify solutions not in the core, such as accounting for bargaining power based on what a participant calculates it contributes to a coalition by joining it (e.g., the Shapley value), or forming coalitions on the basis of no player having a justified objection against any other member of the coalition (e.g., the bargaining set). Also, it is certainly possible to invoke various voting or auctioning protocols to address this problem. We do not criticize the rationale behind these refinements to the basic theory, or the various extra-game-theoretical considerations that may govern the formation of coalitions, such as friendship, habits, fairness, etc. We simply point out that to achieve a reasonable solution it may be necessary to go beyond the strict notion of maximizing individual expectations and employ ancillary assumptions that temper the attitude and behavior of the decision makers

Satisficing negotiation, however, permits controlled degrees of altruism. If agents are willing to lower their standards, as defined by the boldness, b , they may obtain a satisficing compromise, where a joint decision is obtained that is good enough for the group as a whole and good enough for each member of the group. This potential result is guaranteed by the following theorem.

Theorem 1 (*The negotiation theorem.*) *If u_i is individually satisficing for the i th agent, that is, $u_i \in \Sigma_b^i$, then it must be the i th element of some jointly satisficing vector $\mathbf{u} \in \Sigma_b$.*

Proof We will establish the contrapositive, namely, that if u_i is not the i th element of any $\mathbf{u} \in \Sigma_b$, then $u_i \notin \Sigma_b^i$. Without loss of generality, let $i = 1$. By hypothesis, $p_S(u_1, \mathbf{v}) < bp_R(u_1, \mathbf{v})$ for all $\mathbf{v} \in U_2 \times \dots \times U_N$, so $p_{S_1}(u_1) = \sum_{\mathbf{v}} p_S(u_1, \mathbf{v}) < b \sum_{\mathbf{v}} p_R(u_1, \mathbf{v}) = bp_{R_1}(u_1)$, hence $u_1 \notin \Sigma_b^1$. \square

The content of the negotiation theorem is that, under intrinsic satisficing, no one is ever completely frozen out of a deal—every decision maker has, from its own perspective, a seat at the

negotiating table. This is perhaps the weakest condition under which negotiations are possible.

A decision maker possessing a modest degree of altruism would be willing to undergo some degree of self-sacrifice in the interest of others. Such a decision maker may be viewed as an **enlightened liberal**; that is, one who is intent upon pursuing its own self interest but gives some deference to the interests of the group in general. Such a decision maker would be willing to lower its standards, at least somewhat and in a controlled way, if doing so would be of great benefit to others or to the group in general.

The natural way for a decision maker to express a lowering of its standards is to decrease its boldness. Nominally, we may set b_i , the boldness of the i th agent, to unity, which reflects equal weighting of the desire for success and the desire to conserve resources. By decreasing b_i , the agent lowers its standard for success relative to resource consumption, and thereby increases the size of its satisficing set. As $b_i \rightarrow 0$ the standard is lowered to nothing, and eventually every option is satisficing. Consequently, if all decision makers are willing to reduce their standards sufficiently, a compromise can be achieved.

Figure 3 illustrates this negotiatory process. The amount by which b_i must be reduced below unity is a measure of the degree of compromising needed to reach a mutually acceptable solution. As with tension and diversity, however, this degree of compromising is not a measure of performance, but it is a useful figure of merit for assessing the degree of difficulty that is associated with the negotiatory process.

Step 1: Agent i forms $\Sigma_{b_L}^i$ and $\Sigma_{b_i}^i$, $i = 1, \dots, N$; initialize with $b_i = 1$, $b_L = \min\{b_1, \dots, b_N\}$.

Step 2: Agent i forms its compromise set by eliminating all option vectors for which its component is not individually satisficing, resulting in $C_i = \{\mathbf{u} \in \Sigma_{b_L}^i : u_i \in \Sigma_{b_i}^i\}$.

Step 3: Broadcast C_i and b_i to all other participants, receiving similar information from them.

Step 4: Form the satisficing imputation set, $N = \bigcap_{j=1}^N C_j$. If $N = \emptyset$, then decrement b_j , $j = 1, \dots, N$, and repeat previous steps until $N \neq \emptyset$.

Step 5: Agent i implements the i th component of the rational compromise

$$\mathbf{u}^* = \arg \max_{\mathbf{u} \in N} \frac{p_{S_1 \dots S_N}(\mathbf{u})}{p_{R_1 \dots R_N}(\mathbf{u})}.$$

Figure 3: The Enlightened Liberals negotiation algorithm.

This leads to a theory of social behavior that is very different from standard N -person von Neumann-Morgenstern game theory. Whereas, under conventional theory, additional criteria may be required to foster successful negotiations, the sat-

isficing concept builds controlled degrees of compromise into the decision-making procedure. If an agent reaches its limit of compromise before negotiations are successful, it may be forced to declare an impasse, rather than to sacrifice its standards any further.

5. RESOURCE SHARING

The following simple example illustrates the fundamental differences between substantive and intrinsic rationality. Suppose a factory operates N processing stations that function independently of each other, except that, if their power requirements exceed a fixed threshold, they must draw auxiliary power from a common source. Unfortunately, there are only $N - 1$ taps to this auxiliary source, so one of the stations must operate without that extra benefit. Although each station is interested in its individual welfare, it is also interested in the overall welfare of the factory and is not opposed to making a reasonable compromise in the interest of overall corporate success.

Let U denote the set of auxiliary power levels that are feasible for each X_i to tap, and let $f_i: U \rightarrow [0, \infty)$ be an objective function for X_i ; that is, the larger f_i , the more effectively X_i achieves its goal. X_i 's choice is tempered, however, by the total cost of power, as governed by an anti-objective function, $g_i: U \rightarrow [0, \infty)$, such that the smaller g_i , the less the cost. Work cannot begin until all players agree on a way to apportion the auxiliary power. Table 1 displays these quantities for a situation involving three decision makers.

U	f_1	g_1	f_2	g_2	f_3	g_3
0.0	0.50	1.0	0.10	1.0	0.25	1.0
1.0	2.00	2.0	2.00	3.0	0.50	5.0
2.0	3.00	4.0	3.00	6.0	1.00	5.0
3.0	4.00	5.0	4.00	9.0	2.00	5.0

Table 1: The objective functions for the Resource Sharing game.

A standard approach under substantive rationality is to view this as a cooperative game. The payoffs may be obtained by combining the two objective functions, yielding individual payoff functions of, say, the form

$$\pi_i(u_1, u_2, u_3) = \begin{cases} -1 & \text{if } u_j > 0 \forall j \\ \alpha_i f_i(u_i) - \beta_i g_i(u_i) & \text{otherwise} \end{cases},$$

$i = 1, 2, 3$, where α_i , β_i , and μ are chosen to ensure compatible units. To achieve this compatibility, we normalize f_i and g_i to unity by setting $\alpha_i = \frac{1}{\sum_{u \in U} f_i(u)}$ and $\beta_i = \frac{1}{\sum_{u \in U} g_i(u)}$.

The Pareto solution is $\mathbf{u}_P = \{0, 1, 3\}$, but, with an attitude governed by expected utility maximization, X_1 has no incentive to agree to this apportionment. Thus, to solve this problem, a negotiation protocol must be invoked. Of the various protocols that are possible, the only one that does not require assumptions

additional to that of self-interested expectations maximization is the core. Unfortunately, the core is empty for this game. Essentially, this is because only two decision makers can share in the auxiliary power source, effectively disenfranchising the third decision maker. This situation potentially leads to an unending round of recontracting, where participants continually make offers and counter offers in a fruitless attempt for all to maximize their expectations.

Let us now view the decision makers in their true character as enlightened liberals who are willing to accept solutions that are serviceably good enough for both the group and the individuals. From the point of view of the group, an option is satisficing the joint selectability exceeds the joint rejectability scaled by boldness. We define joint rejectability as the normalized product of the individual costs functions, namely,

$$p_{R_1 R_2 R_3}(u_1, u_2, u_3) \propto g_1(u_1)g_2(u_2)g_3(u_3),$$

where “ \propto ” means the function has been normalized to sum to unity. To compute the joint selectability, we note that, under the constraints of the problem, only two of the agents may use the auxiliary power source. We may express this constraint by defining the joint selectability function as

$$p_{S_1 S_2 S_3}(u_1, u_2, u_3) \propto \begin{cases} p_{S_1}(u_1)p_{S_2}(u_2)p_{S_3}(u_3) & \text{if } \mathbf{u} \in \Pi \\ 0 & \text{otherwise} \end{cases}$$

where Π is the set of all triples $\mathbf{u} = \{u_1, u_2, u_3\}$ such that exactly one of the entries is zero. The individual rejectability and selectability marginal mass functions are obtained by summing over these joint mass functions according to the rules of probability theory.

The enlightened liberals algorithm yields, for $b > 0.8$, an empty satisficing imputation set. But, when b is decremented to 0.8, the satisficing imputation set is $\mathbf{N} = \{\{0, 1, 3\}, \{0, 2, 3\}, \{0, 3, 3\}\}$ and the rational compromise is $\mathbf{u}^* = \{0, 1, 3\}$ which, coincidentally, is the Pareto optimal solution. It is not surprising that, at unity boldness, there are no options that are simultaneously jointly and individually satisficing for all participants, since there is a conflict of interest (recall that the core is empty). But, if each individual adopts the point of view offered by intrinsic rationality, it gradually lowers its personal standards to a point where it is willing to be content with reduced benefit, provided its costs are reduced commensurately, in the interest of the group achieving a collective goal. The amount b must be reduced to reach a jointly satisficing solution is an indication of the difficulty experienced by the participants as they attempt to resolve their conflicts. Reducing boldness is a gradual mechanism for decision makers to subordinate individual interest to group interest. This mechanism is very natural in the regime of making acceptable tradeoffs, but is quite foreign to the concept of maximizing expectations (“you get what you pay for” versus “nothing but the best”).

The diversity and tension values for this decision problem are given in Table 2. We interpret these values as follows.

Agent	Diversity	Tension
X_1	0.55	0.93
X_2	0.03	1.30
X_3	1.21	0.73
Group	2.85	0.51

Table 2: Diversity and Tension for Resource Sharing Game.

Group diversity is high and group tension is low, indicating that, as a group, the system is fairly well suited for its environment, and that the system is powerful enough to make good decisions. Individually, X_2 has the lowest diversity and the highest tension. This situation is reflected in the structure of N , where we see that X_2 has several choices that are good enough, but is either dubious or ambivalent about all of them. Thus, X_2 experiences the most conflict in making decisions. X_3 is quite content with its decision and so is X_1 . The fact that X_1 is not conflicted as measured by diversity and tenseness may appear somewhat contradictory, since it is X_1 who ends up sacrificing for the benefit of the group. But these figures of merit are not intended to be metrics of performance, only of the intellectual power of the decision maker, in terms of its conflict between selectability and rejectability.

6. CONCLUSION

An intelligent agent is, first and foremost, a decision maker, regardless of the problem context, the way knowledge is represented, or the criteria used to define performance. One way to assess the functionality of the agent is to provide it with a means to evaluate introspectively its own fitness, or suitability, to function in its environment. Satisficing decision theory provides this capability. Although the figures of merit associated with these fitness evaluations are not measures of performance, they are useful measures of the innate intellectual (decision-making) power of the agent.

References

- [1] M. Bazerman. A critical look at the rationality of negotiator judgement. *Behavioral Science*, 27:211–228, 1983.
- [2] M. H. Bazerman and M. A. Neale. Negotiator rationality and negotiator cognition: The interactive roles of prescriptive and descriptive research. In P. H. Young, editor, *Negotiation Analysis*, pages 109–129. Univ. of Michigan Press, Ann Arbor, MI, 1992.
- [3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, New York, 1991.
- [4] R. D. Luce and H. Raiffa. *Games and Decisions*. John Wiley, New York, 1957.

- [5] A. Rapoport and C. Orwant. Experimental games: a review. *Behavioral Science*, 7:1–36, 1962.
- [6] H. A. Simon. A behavioral model of rational choice. *Quart. J. Econ.*, 59:99–118, 1955.
- [7] H. A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138, 1956.
- [8] H. A. Simon. Rationality in psychology and economics. In R. M. Hogarth and M. W. Reder, editors, *Rational Choice*. Univ. Chicago Press, Chicago, 1986.
- [9] W. C. Stirling and M. A. Goodrich. Satisficing games. *Information Sciences*, 114:255–280, March 1999.
- [10] W. C. Stirling, M. A. Goodrich, and D. J. Packard. Satisficing equilibria: A non-classical approach to games and decisions. *Autonomous Agents and Multi-Agent Systems Journal*, 2000. To appear.

What is the Value of Intelligence?

Thomas Whalen

Professor of Decision Sciences, Department of Management

Robinson College of Business, Georgia State University

whalen@gsu.edu

Abstract

Probably the most widespread and significant existing “performance metric for intelligent systems” is the dollar premiums that employers are willing to pay to recruit and retain more intelligent human employees compared to less intelligent ones. This paper examines some of the aspects driving this economic metric in the search for analogies that may be useful in establishing performance metrics for constructed intelligent systems. Aspects considered include Language Understanding & Capacity to Act, Goal-Directedness, Autonomy and Unpredictability, Information, Uncertainty, World Models, and Self-Models and Self Awareness. The paper concludes with a discussion of performance metrics for human intelligence and a brief prospectus for the role of economic considerations in assessing the Vector of Intelligence

Keywords: *economic value, intelligence*

1. Introduction

Much of the discussion leading up to the conference on “Performance Metrics for Intelligent Systems” focuses on an “inner” view of intelligent performance, or rather of intelligence itself. This inner view takes two very different forms: components like memory or MIPS that must be present inside an intelligent system, and metaphysical questions about the “inner life” of an intelligent system, such as questions of consciousness.

Rather than try directly to add to this interesting and valuable train of thought, this paper approaches the subject of performance metrics for intelligent systems from an external perspective. The question under consideration here is “What is the economic value of intelligence?” Most of the discussion will concern the market value of human intelligence, in order to look for useful analogies for understanding and measuring the economic value of intelligence in constructed systems.

Individuals treasure intelligence in themselves and their friends and family for a variety of reasons, most of which lead rapidly into the spiritual or metaphysical realm, or, if you prefer, into the most complex challenges of sociobiology. Either way, creating a “performance metric” for intelligence in this context seems neither feasible nor especially desirable.

On the other hand, consider the owners of a medium-sized business, who need to hire a number of employees to perform various tasks in the firm. Why should the owners pay a higher salary and go through a more difficult and expensive recruitment process to hire a more intelligent employee when they can get a less intelligent employee with the same training and experience more cheaply? To the extent we can give a quantitative answer to this question, the dollar premium a business is willing to pay for intelligence is a financial “performance metric for intelligent employees” within the context of the job at hand. Understanding how these dollar premiums arise in a variety of employment situations can give important clues on how to put a value or “metric” on the performance of intelligent machines.

There are three distinguishable ways in which a smarter employee can be worth more money to a business than a stupid one with equivalent training and experience. These are: doing what I say, doing what I want, and doing what I need.

2. Language Understanding & Capacity to Act

At the most fundamental level, “**do what I say,**” an intelligent laborer can follow instructions better than a stupid laborer. Smart employees can follow instructions that are more complex, less detailed, and require less time and effort (in other words, less money) to prepare. Since they are less apt to misunderstand instructions, they require less money to be spent on supervising them than is the case for less intelligent employees with equal motivation. For constructed systems, the equivalent is an expressive command language; one that is the “natural language” for describing the task at hand, whether it resembles a spoken human language, a specialized technical language, or a graphical interface. Allied with this, of course, is the capacity to actually carry out the instructions, which some have referred to as the “body” as opposed to the “mind” of the intelligent constructed system.

3. Goal-Directedness

It is possible to view the next level, “**do what I want,**” as simply an elaboration of the ability of smarter employees to follow instructions that are less detailed. However, businesses look hard for intelligent skilled craftsmen who can be told what goals to accomplish without needing to be told how to do so, and reward them with higher wages and better

treatment. A major topic of discussion has been the role of **goal - directedness** in intelligent systems. In the world of human employment, a laborer (first level) is given instructions about how to do a job; the goal may be implicit in the instructions but is not an integral part of them from the laborer's point of view. A craftsman (second level), on the other hand, takes the goals provided by the employer and carries them out without further instruction. To do this, the craftsman needs experience and training, but also puts more intelligence into the work than the laborer does.¹

Over time, a job may become more routinized, so that what originally required highly intelligent goal-seeking behavior later requires only the following of rote instructions. This can occur at either the structural level as the instructions are written down for others, or within an individual as long experience with a job eventually allows it to be done “without thinking.” The equivalent to this process in the area of constructed systems would be the replacement of complex, “intelligent” processes of sophisticated search and behavior generation with stereotyped program modules or hardware gadgets, reducing the “intelligence” used by a constructed system while maintaining or even enhancing its performance.

4. Autonomy and Unpredictability

At both of the first two levels, management wants behavior of the employee to be predictable. Intelligence means autonomy in the sense that, given equivalent training and motivation, the intelligent employee does what is expected of him or her without close supervision while the stupider employee in the same job needs to be watched all the time. However, autonomy in this context is almost the opposite of creativity, spontaneity, or unpredictability; it is the stupid employee, not the smart one, who comes up with the most surprises.

It is only at the highest level, “**do what I need,**” that businesses value unpredictability in their employees and consultants. Even here, there are two degrees of unpredictability. Most of the time a person or company seeks advice on matters of law, engineering, medicine, or other fields, the advice has no “information” value if the one requesting it already knew the answer; nevertheless, routine advice needs to be in line with professional standards. For example, though I do not want to be able to predict what my personal physician is going to tell me, I want it to be essentially the same as what any competent physician would say given the same knowledge about me; in other words, I want my physician's behavior to be essentially predictable by other physicians. It is only

if I am suffering from an extremely serious disease, or if I am knowingly participating in a clinical experiment, that I want my physician to do something that will surprise the medical profession!

5. Information

Some of the discussion about performance metrics for intelligent systems has debated the applicability of entropy or other aspects of information theory to measuring intelligence. Fundamentally, “Information” implies informing somebody about something they didn't already know. From this point of view, an employer wants a laborer's work to provide no new information output at all, but a more intelligent laborer requires less information input than an unintelligent one. A craftsman working at the second level of “doing what I want” takes compact information about goals rather than lengthy information about procedures; the craftsman's work in sense generates “information” to the employer about the methods used, but this is information that normally is of no great interest to the employer. It is only at the highest level, that of the professional employee, that the employer is concerned about receiving information output from the employee.

		Information Input	Information Output
Laborer	Do what I say	High, procedural	Ideally none
Craftsman	Do what I want	Low, goal-oriented	Uninteresting
Professional	Do what I need	Various	Essential

6. Uncertainty

The more uncertain the job environment is, the more valuable an intelligent employee becomes. Procedural instructions about an uncertain job environment must become a complex collection of “ifs” and branches, compared to a more linear set of instructions for a job in a less uncertain environment. Businesses have to pay more for employees intelligent enough to follow such complex instructions than they do for employees whose jobs do not contain much uncertainty.

For sufficiently high levels of uncertainty in the job environment, management finds it unprofitable to prepare procedural instructions in a form that even the smartest laborer can follow. Instead, it is more economical to hire craftsmen who only need to be told the employer's goals and essentially left to implement those goals according to their own skills and

¹ Note that my focus here is on the degree of intelligence demanded by the job, not on the intelligence possessed by the human being doing it. Job demands place only a lower bound on the worker's intelligence. Nevertheless, the more intelligence the job demands, the more the performance of an intelligent employee will overshadow that of a less intelligent one.

intelligence. The fundamental problem with the “Chinese Room” thought experiment is that, while it might in principle be possible to prepare and index a set of stimulus-response instructions so extensive as to allow the occupant of the room to carry on a conversation in Chinese without any knowledge of the language, it is in fact such an immense task that it would be far cheaper and easier to build a machine that actually understood Chinese (and easier still to hire a human who understands Chinese to sit in the room!).

At the highest levels of uncertainty (or extreme complexity, which as Zadeh points out has many of the same effects) management can no longer be sure what goals are feasible or profitable, and so seeks expensive and potentially surprising guidance from professionals, and perhaps some day from constructed systems that produce “useful surprises” at a professional level.

7. World Models

It is very rare for an employer to ask about an employee’s internal model of the world or to pay a higher salary on account of it. Laborers are paid to follow instructions intelligently in the real world, and craftsmen are paid to ply their trades intelligently in the real world. Whether or not they use an internal model of the world to do so is of no economic importance except as it is reflected, at one or more removes, in their performance.

Professionals are paid to give “useful surprises” to their employers or clients. This information (and actions informed by it) generally have to do with the real world, though at times professionals may be asked for opinions about hypothetical situations. Even then, usually it is irrelevant whether the answer comes from stored knowledge, experimentation, or the exercise of a simulation-like model in the professional expert’s head. The exception is when the professional is explicitly asked to provide a model, but in that case the model is no longer an internal one, but an external analogy, flow-chart, or computer simulation.

8. Self-Models and Self Awareness

Certainly, all of a firm’s (human) employees have a self-model, a self-awareness, a consciousness. But only in a few “helping professions” such as psychiatry or the clergy is an above-average endowment in this area considered an advantage to job performance. Employers value some limited facets related to self-awareness such as taking pride in one’s work and being safety-conscious, but outstanding self-consciousness and self-absorption are not considered signs of outstandingly valuable intelligence by employers. Thus, with regard to constructed systems, it might be an economically important goal to build machines that “care” about doing a good job and know how to take care of themselves and those around them. But we should not insist on a robotic Mother

Teresa; it would be a magnificent achievement to create a working system that was as caring and careful as a seeing-eye dog.

9. Performance Metrics

Unlike constructed systems, human employees cannot be opened up to inspect their components. Thus, employers in search of intelligent employees rely on a variety of benchmark tasks. Occasionally, they may use a benchmark task that tries to screen out the effects of knowledge to focus on pure intelligence -- examples include IQ tests and programmer aptitude tests. However, since job performance is more important than what mix of knowledge, intelligence, and other endowments it arises from, most benchmark tasks measure performance without much concern about the mix. The most common benchmark task is performance on similar jobs in the past.

Another interesting benchmark is formal education. Completing any program of study implies an ensemble of intelligence, knowledge, and skills for learning, writing, and simply sticking to a task. The education most valued by employers adds to this a body of knowledge relevant to the job. However, for complex and unpredictable environments, it may not be possible to specify in advance what body of knowledge will be required. In such a case, a broad “general education” demonstrates that a person has an advanced ability, refined by varied practice, to learn whatever is required in a new situation. With respect to constructed systems, a design team that hones and demonstrates their product’s ability to learn and excel in a wide variety of problem environments, including artificial ones as well as real ones, can command a higher price for their machines than a design team that only trains their system on what is “relevant” to its expected tasks, at least from customers whose jobs are at the high end of uncertainty or complexity.

Performance metrics for intelligent systems based on board games like chess and backgammon or parlour games like the Turing test can be very useful in addressing philosophical questions about what it means to be intelligent, and technological questions about how to implement it, but they are of little direct economic interest. In particular, to pass the Turing test in a job application context, an intelligent system would have to refrain from showing any levels of ability not common among humans, and also to demand the same levels of salary and benefits as a human. What is needed, instead, is a set of benchmark tasks, probably job-specific, with one or more of the following characteristics:

- Instructions are so complicated that it is more profitable to seek an intelligent laborer system that understands them, than to seek an unintelligent “Chinese room”

type system to follow the instructions without understanding.

- The environment is so complicated and uncertain that it is more profitable to seek an intelligent craftsman system that accepts exogenous goals and carries them out according to its own skills and intelligence, rather than to seek an unintelligent system that simply follows instructions.
- The situation is so fuzzy that it is more profitable to seek an intelligent professional system to determine what goals are appropriate (presumably given exogenous meta-goals) and do surprising things for the benefit of the organization, rather than to seek an unintelligent system that simply and predictably carries out exogenous goals

To be useful, an intelligent constructed system must provide a better cost/benefit ratio than any combination of human being(s) and unintelligent constructed system(s). If more than one intelligent constructed system meets this test,

then the one with the best cost/benefit ratio, not necessarily the smartest one, will be chosen.

10. Economics and the Vector of Intelligence

The “white paper” for the 2000 Conference on Performance Metrics for Intelligent Systems lists 25 potential coordinates for a possible Vector of Intelligence. A major challenge is to find ways to systematically quantify or otherwise specify the values of these “coordinates.” Without detracting from the usefulness of methods oriented toward philosophy of mind, toward control engineering, or toward academic computer science, let me propose an economic approach to measuring each of the 25 coordinates summarized in the following table. In this economic approach, the challenge would be to estimate the derivatives of system cost/benefit ratio in a benchmark problem to “memory temporal depth,” “number of objects that can be stored,” ... et cetera. The second derivative is as important as the first since most or all of these coordinates are subject to diminishing or even negative returns.

Twenty-Five Potential Coordinates for the Vector of Intelligence (from the White Paper)

- (a) memory temporal depth
- (b) number of objects that can be stored
- (c) number of levels of granularity in the system of representation
- (d) the vicinity of associative links taken in account during reasoning of a situation, or
- (e) the density of associative links
- (f) the vicinity of the object in which the linkages are assigned and stored (associative depth)
- (g) the diameter of associations ball (circle)
- (h) the ability to assign the optimum depth of associations
- (i) the horizon of planning at each level of resolution
- (j) the horizon of extrapolation at a level of resolution
- (k) the response time
- (l) the size of the spatial scope of attention
- (m) the depth of details taken in account during the processes of recognition at a single level of resolution
- (n) the number of levels of resolution that should be taken into account during the processes of recognition
- (o) the ratio between the scales of adjacent and consecutive levels of resolution
- (p) the size of the scope in the most rough scale
and the minimum distinguishable unit in the most accurate (high resolution) scale
- (q) an ability of problem solving intelligence to adjust its multi-scale organization to the hereditary hierarchy of the system,
- (r) dimensionality of the problem (the number of variables to be taken in account)
- (s) accuracy of the variables
- (t) coherence of the representation constructed upon these variables
- (u) limit on the quantity of texts available for the problem solver for extracting description of the system 20
- (v) frequency of sampling and the dimensionality of the vector of sampling
- (w) cost-functions (cost-functionals)
- (x) constraints upon all parameters
- (y) cost-function of solving the problem

On the Computational Measurement of Intelligence Factors

José Hernández-Orallo

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, Camí de Vera s/n, 46022 València, Spain
E-mail: jorallo@dsic.upv.es.

Abstract. In this paper we develop a computational framework for the measurement of different factors or abilities usually found in intelligent behaviours. For this, we first develop a scale for measuring the complexity of an instance of a problem, depending on the descriptive complexity (Levin *LT* variant) of the ‘explanation’ of the answer to the problem. We centre on the establishment of either deductive and inductive abilities, and we show that their evaluation settings are special cases of the general framework. Some classical dependencies between them are shown and a way to separate these dependencies is developed. Finally, some variants of the previous factors and other possible ones to be taken into account are discussed. In the end, the application of these measurements for the evaluation of AI progress is discussed.

1 Introduction

Are AI systems of today more intelligent than those of 40 years ago? Probably the answer is a clear yes, at least for some of the current systems. However, another different question is ‘How much more intelligent?’, and, even more, in which aspects are they more intelligent?

In this paper we investigate a framework for the evaluation of such a progress in different factors, extending in a natural way the work endeavoured in [12] and [11], specific for only some inductive factors. For such an extension, the main aim should be to develop the less number of factors as possible, by proposing general factors instead of specific ones. Moreover, the framework would allow to studying their theoretical correlations, and reducing, when possible, a factor to another. This leads finally to a group of tests that can be adapted and implemented for measuring different abilities of AI systems.

First of all, we must ascertain three problems for any evaluation of the ability of solving a problem: to give a general scale of a complexity of the problem, to settle the unquestionability of the solution to the problem and to establish a way to know whether the subject has arrived to the solution.

Computational complexity scales problems according to the time different kinds of machines require to solve them in the general case by using the optimal algorithm possible. However, most problems of interest in AI are NP-complete. But, remarkably, some instances of NP-complete problems are easier than instances of polynomial problems. This assertion seems to be contradictory, since any instance has an algorithm to solve that instance in linear or even constant time (the program “if the input is x print the solution y ”), so there is apparently no reason for stating that an instance can be easier than another. This has been shown to be false up to an extent, because for some problems it is better (shorter) to give a more general solution than the specific solution for an instance of the problem. This has been formalised under the notion of “instance complexity” (see e.g. [16]), which gives the shortest solution to an instance of a problem provided it does not give a contradictory solution for other instances of the same problem.

However, instance complexity is only of interest for large instances of a considerable descriptive complexity (or for sets of instances). Moreover, the difficulty of the problem is not usually related to the descriptive complexity of the solution. For instance, the descriptive complexity of the answers given by a theorem prover (an acceptor) are very short, namely one bit to say

‘yes’ or ‘no’. In the same way, the hardness of a prediction problem cannot be measured by the descriptonal complexity of the element predicted, but rather by the complexity of the reason why the element has been predicted. The idea is then to measure the descriptonal complexity of the ‘justification’ or ‘explanation’ of the solution. Consequently, any cognitive skill can be measured within this framework provided that problem and solution can be formalised computationally.

The paper is organised as follows. After Section 2, where some notation is introduced, Section 3 gives a general formula of the hardness of the instance of a problem, by clarifying how to generalise the concept of ‘explanation’ of a solution to a problem. Section 4 addresses the issue of specialising it for deductive abilities and discusses their measurement. Section 5 does the same thing for inductive abilities, but recognising that it is necessary to solve the unquestionability problem. Section 6 deals with their dependencies and the possibility of taking other factors into account. Section 7 discusses the applications of these measurements, especially for the evaluation of automated reasoning and machine learning systems. Section 8 closes the paper with the results and open problems.

2 Preliminaries

Let us choose any finite alphabet Σ composed of symbols (if not specified, $\Sigma = \{0,1\}$). A string or object is any element from Σ^* , with \circ being the composition operator, usually omitted. By $\langle a, b \rangle$ we denote a standard recursive bijective encoding of a and b , such that there is a one-to-one correspondence between $\langle a, b \rangle$ and each pair (a, b) . Note that this usually takes more bits than $a \circ b$. The empty string is denoted by ε . The term $l(x)$ denotes the length or size of x in bits and $\log n$ will always denote the binary logarithm of n .

The complexity of an object can be measured in many ways, one of them being its degree of randomness [14], which turns out to be equal to the shortest description of it. Descriptonal Complexity, Algorithmic Complexity or Kolmogorov Complexity was independently introduced by Solomonoff, Kolmogorov and Chaitin to formalise this idea, and it has been gradually recognised as a key issue in statistics, computer science, AI and cognitive science [16][6].

The Kolmogorov Complexity of an object, defined as the shortest description for it, usually denoted by C (plain complexity) or K (prefix-free complexity) turns out to be not computable in general, due to the halting problem. One solution for this is to incorporate time in the definition of Kolmogorov Complexity. The most appropriate way to weight space and time execution of a program, the formula $LT_\beta(p_x) = l(p_x) + \log \tau_\beta(p_x)$, where τ is the number of steps the machine β has taken until x is printed by p_y , was introduced by Levin in the seventies (see e.g. [15]). Intuitively, every algorithm must invest some effort either in time or demanding/essaying new information, in a relation which approximates the function LT . The corresponding complexity, denoted by Kt (see e.g. [16]) is a very practical alternative to K .

3 Problem Complexity by Its Explanation Complexity

Consider a problem instance π as a tuple $\langle S, C, I, A, \phi \rangle$ where S is the context or working system where the problem can be established, C is a Boolean function which represents a (syntactical) validity criterion, I is the presentation of the instance, A_i is the answer and ϕ is a (semantical) verifier¹. The general problem is denoted by $\pi(\cdot)$ as the tuple $\langle S, C, \phi \rangle$.

We say that E is an explanation for the problem instance π iff E is valid, i.e. $C(\langle S, I, E \rangle) = true$, and E is a means to obtain the solution, i.e., $\phi(\langle S, I, E \rangle) = A_i$.

From here, it is easy to adapt the definition of Kt to measure the hardness of a problem. Namely, the hardness of a problem instance $\pi \langle S, C, I, A, \phi \rangle$ is then defined as:

$$H(\pi) = \min\{LT(E|\langle S, C, I \rangle) : E \text{ is an explanation for } \pi\} \quad (1)$$

¹ Both C and ϕ could be joined in one function. We have preferred to separate them, because later it will be useful to distinguish between both parts of a correct solution, in order to establish purer factors.

For instance, the hardness of a search problem is usually estimated by the size of the search space. If the search problem is complex, it is necessary to say which branches have been selected in order to arrive to the solution, or either a long time is necessary to explore (and make backtracking) to the misleading ones. It is the function LT which finds a compromise between the information which is needed to guide the search and the logarithm of the time that is also needed to essay all the branches. On the other hand, if the search problem is linear (one possible branch), it is very easier to describe the problem (just follow the rules in the only possible way). However, for very long derivations, the inclusion of time can make hardness high too.

For the evaluation of a subject's ability of solving a kind of problem $\pi(\cdot)$ it is necessary to generate a set of instances of that problem of different hardness. In order to scale the instances more properly, we introduce the concept of k -solvability. An instance of a problem $\pi = \langle S, C, I, A, \phi \rangle$ is k -solvable iff k is the least positive integer number such that:

$$H(\pi) \leq k \cdot \log l(I) \tag{2}$$

The use of $\log l(I)$ is justified by the fact that, once the general problem is known, each instance must be 'read' and this takes at least $l(I)$ steps.

Once given a general scale of a complexity of the problem, it is then easy to make a test from the previous definition, provided that the unquestionability of the solution to the problem is clear. Unquestionability can only be addressed depending on the kind of problem (we will see this for deductive abilities and especially for inductive abilities in the following sections). Finally, there is no way to know whether the subject has arrived to the solution if the explanation is not given (and usually the explanation is difficult to check or the subject may not be able to express the explanation in a comprehensible form). For instance, the subject may have given the right solution but maybe due to wrong derivations. Fortunately, in the case of multiple solutions, this situation will be discardable in the global reckoning of the test. In the case of few solutions, such as 'yes'/'no', it is then necessary to penalise the errors by using some formula that takes into account the possibility of guessing the right answer 'by error'.

Another question is the time limit for making the test. This would highly depend on the factor to be measured, and whether there is a special interest on evaluating the ability to solve a given problem or the ability to solve it quickly. The selection of the time limit and the evaluation of the score according to it could be very interesting for evaluating resource-bounded rational systems.

Finally, we have not considered the possibility of multiple correct explanations for the same solution, which would suggest a modification of (1). Consider the situation of the best explanation with $LT = n$, but several other explanations of $LT = n + 1$. Intuitively, the existence of these other explanations also affects the easiness of the solution. However, this is very difficult to evaluate in practice because there are always infinite slight variations of the best explanation (void steps, redundancies, etc.), so the previous situation is extremely frequent (if not inevitable). It is then assumed that for every k :

$$\begin{aligned} \text{card}\{ E : LT(E) = k \text{ and } C(\langle S, I, E \rangle) = \text{true} \text{ and } \phi(\langle S, I, E \rangle) = A_i \} &<< \\ \text{card}\{ E : LT(E) = k \text{ and } C(\langle S, I, E \rangle) = \text{true} \} & \end{aligned} \tag{3}$$

In other words, we assume that the proportion of valid and correct explanations wrt. valid explanations is very small.

Once a general framework is established, let us study which deductive and inductive abilities are feasible and interesting to be measured within it.

4 Deductive Abilities

Apparently, deductive abilities are much easier to measure, because there is no possible subjectivity in the correct answer; given the premises and the way to operate with them, only one answer is possible.

An instance of a deductive problem $\pi = \langle S, C, I, A, \phi \rangle$ can be defined in terms of the previous framework in the following way: S corresponds to the set of axioms or axiomatic system, C is a Boolean function which says what is a valid application of the axioms, I is the instance of the deductive problem, A_i the answer and ϕ is a verifier, i.e., $\phi(\langle S, I, E \rangle) = A_i$, in this case, a verifier that checks whether A_i is a result of applying a solution to I in S .

In this case the explanation E is represented by a *proof* in S stating that A_i is a the result of I or, in other words, a derivation from I to A_i .

Example: Consider for instance an accepter that tells whether a proposition is a theorem or not. Let S be the axioms of arithmetic. Let C a function that tells that a derivation is valid according to the rules of application of the axioms, and let I be the instance “Is Fermat’s famous conjecture true?” (recently a theorem). Which is the hardness of the solution $A = \text{‘yes’}$? The descriptonal complexity of A (which is just yes) would say that the instance is very easy, however its hardness given by H turns out to be the LT of the proof with less LT . Consider instead the instance “solve $2+3$ ” which, also with a low complexity of $A = 5$, turns out to be simple, because the derivation is describable easily and shortly from $\langle S, C, I \rangle$. In general, any calculation is shortly describable, so its hardness will depend solely on its temporal cost.

According to this example, we can distinguish some classical deductive problems that can be measured. In particular, the following factors are distinguished:

- **Calculus Ability:** in this special case, C only allows a specific and deterministic application of the rules or axioms of S . In this case the search space is linear. As it has been said before, its complexity is exclusively given by the logarithm of the time which is needed from the input I to the output A_i . This ability is not of much interest to be measured nowadays, since it is better done by computers than humans, and it would finally measure the computational power of the subject / machine.
- **Derivational Ability:** in this case, C only allows a varied application of the rules or axioms of S . Consequently, the search space is open. The complexity is then given by a compromise between the logarithm of the time which is needed to know that a branch leads to no solution, and some information that may say which branches to take (and which ones not to take).
- **Accepter Ability (proving ability):** It is a special case of the previous ability, with the special feature that I can only be ‘yes’ or ‘no’. Theoretically, there is no reason for expecting that a subject has a different result in this problem that in the previous one.

The way to implement a concrete test for the previous ability is not complicated. For calculus ability, it is just necessary to generate some derivations. Their length will determine the time which is needed to follow them. On the contrary, for the other two abilities, it is necessary to generate a possible derivation, and look that there are no shorter equivalent derivations. This, in general, will be extremely costly, growing exponentially according to the value of k -solvability. Fortunately, there is no need for efficiency here. A hard test can be generated during days, even weeks, and then passed to several subjects.

5 Inductive Abilities

A sequential inductive problem $\pi = \langle S, C, I, A, \phi \rangle$ can also be defined in terms of the previous framework in the following way: S corresponds to the background knowledge, I is a sequential evidence (with $l(I) = n$), C is a Boolean function which represents the hypothesis selection criterion (e.g. simplicity), A_i is the prediction of the $(n + 1)$ th element of the sequence and ϕ is a verifier, i.e., $\phi(\langle S, I, E \rangle) = A_i$, in this case, a verifier that checks whether A_i is the $(n + 1)$ th element given by the hypothesis with the background knowledge S and also checks whether both cover I .

In this case the explanation E is represented by a ‘hypothesis’ wrt. S that affirms that A_i is ‘what follows’ I or, in other words, a prediction from I .

Example: Consider for instance a prediction problem. Let S be a background knowledge, containing, among other things, the order of the Latin alphabet. Let C a function that tells that a hypothesis is

good according to a selection criterion, and let I the instance “aaabbbccdddeefffgggh”. Which is the hardness of the solution $A_i = 'h'$? The descriptonal complexity (in LT terms) of the hypothesis is again what is taken into account.

The main question of evaluation of induction is that of inquestionability. Even if the selection criterion is given, two plausible explanations may differ slightly, and the selection criterion would give that one is slightly better than the other, but this would depend highly on the descriptonal mechanism used. In [12] and [11] this difficult problem is addressed, according to a comprehensive criterion, a variant of the simplicity criterion based on Kolmogorov Complexity in the style of Solomonoff [19], but ensuring that the data is covered comprehensively, i.e. without exceptions. Accordingly, the *simplest explanatory description*, denoted by $SED(x|y)$, is defined in [11] as the simplest (in LT terms) description which is comprehensive wrt. the data x given the background knowledge y . To ensure unquestionability, the examples are selected such that there are no alternative descriptions of similar complexity that give a different description. Finally, there is a small possibility that a good prediction is given by a ‘wrong’ explanation. This probability may be neglected in the tests or corrected by a penalising factor in the score of wrong results.

From here, partially independent factors can be measured by using extensions of the previous framework. For instance, inductive abilities, such as sequential prediction ability, knowledge applicability, contextualisation and knowledge construction ability can be measured in the following way:

- Sequential Prediction Ability: several unquestionable sequences of different k -solvability are generated. A test for this ability has been generated in [12] and passed to humans, jointly with a typical psychometrical test of intelligence. The correlation showed that this is one of the fundamental factors of intelligence, although more experimentation is to be done.
- Inductive Knowledge Applicability (or ‘crystallized intelligence’): a background knowledge B and a set of unquestionable (with or without B , denoted by $H(x_i|B)$ and $H(x_i)$ respectively) sequences x_i are provided such that $H(x_i|B) = H(x_i) - u$ but still $SED(x_i|B) = SED(x_i)$. The difference of performance between cases with B and without B is recorded. This test would actually measure the application of the background knowledge depending on two parameters: the complexity of B and the usefulness of B , measured by u .
- Inductive Contextualisation: it is measured similarly as knowledge applicability but supplying different contexts B_1, B_2, \dots, B_T with different sequences $x_{i,t}$ such that $H(x_{i,t}|B_t) = H(x_{i,t}) - u$. This multiplicity of background knowledge (a new parameter T) distinguishes this factor from the previous one.
- Inductive Knowledge Construction (or learning from precedents): a set of sequences x_i is provided such that there exists a common knowledge or context B and a constant u such that for $H(x_i|B) \leq H(x_i) - u$. A significant increase of performance must take place between the first sequence and the later sequences. The parameters are the same as the first case, the complexity of B and the constant u .

It is obvious that these four factors should correlate, especially with the first one, which constitutes a necessary condition for having a minimal score in the other factors.

6 Dependencies and Other Factors

Although there is a common (but arguable) view of induction and deduction as inverse processes, they are not inverse in the way they use computational resources. In fact, any inductive process requires deduction to check the hypotheses, thus, obviously, inductive ability is influenced by deductive ability. This has been usually recognised by IQ tests, where deductive and inductive abilities usually correlate. Due to this fact, inductive factors usually are the main part of intelligence tests, because deductive abilities are implicitly evaluated.

However, if we are looking for ‘pure’ factors the question is whether there is a way to separate this deductive ‘contamination’ in inductive factors.

The idea is to provide ‘external’ deductive abilities when measuring inductive factors, in order to ‘discount’ the deductive effort than otherwise should be done. For this, given a problem $\pi = \langle S, C, \phi \rangle$ it is only necessary to provide an ‘oracle’ which computes ϕ in constant time. The subject must only guess models (hypotheses) and check them in the oracle, by providing the hypothesis to it and comparing the results with the evidence I . This would measure the ‘creative’ part of induction. In the following, let us denote by ‘purely’ inductive the corresponding factors to those highlighted in the previous section which result from providing the oracle.

This resembles a ‘trial and error’ problem considering reality acting as the oracle. The issue is how to implement this in a feasible way, especially for evaluating complex agents or even human beings. The best way, in our opinion, is the construction of a ‘virtual’ world where the subject to be evaluated can interact and essay its hypotheses with no effort.

In a similar way as the oracle for ϕ , some difference could be estimated if the syntactical machine C is (also) given. Although this would not be much representative for deduction, for induction it would discount the ability of working with the selection criterion, which is an important trait of induction.

Nonetheless, deductive ability is also influenced by inductive ability as long as the problems become harder. Some lemmata or rules can be generated by an intelligent subject in order to help to shorten the proof from the premises to the conclusion. This may explain why artificial problem solvers without inductive abilities have not been able to solve complex problems, and this is especially clear in Automatic Theorem Proving. Consequently, recent systems are beginning to use ML techniques for improving performance. Background knowledge could also be examined in deduction, provided S includes the axioms but also some useful properties. This finally gives similar factors as those given for induction:

- Deductive Knowledge Applicability: how lemmata or properties are used for a deductive problem.
- Deductive Contextualisation: the ability of using different contexts for different problems.
- Deductive Knowledge Construction: this will measure the increase of performance between first instances and last ones.

Finally, we have given a measurement for sequential induction, and it seems interesting to evaluate non-sequential induction as well, where an unordered set of elements is given as evidence from an unknown function that maps whether an element belongs to a set. In this case, the test could give some possible values which might be members of the set, although only one of them is really in it. Solomonoff formalised deterministic (sequential) prediction [19] and recently, has formalised non-sequential prediction [21]. This problem is similar to the inductive problem of learning a Boolean classifier and can be extended to the case of a general classifier. To eliminate the deductive contamination of the measurement of non-sequential induction, the ‘oracle’ ϕ should be a classifier, telling, given a hypothesis, to which class the element belongs. The essay of an ‘oracle’ that accepts several elements at a time should be considered as well.

Once the basic deductive and inductive factors have been recognised, the question is whether there are many other factors which are relevant to be measured. For instance, memory or ‘memoisation ability’ is a factor that is knowledge-independent and it can be easily measured. However, this factor is not very interesting for AI nowadays.

Other factors, such as analogical and abductive abilities can be shown to be closely connected to inductive and deductive abilities both theoretically and experimentally. A first approach for measuring them has been attempted in [12], and the test applied to human beings has shown the correlation with inductive abilities.

However, not every factor is meaningful. Factors like “playing chess well” are much too specific to be robust to the subject’s background knowledge. However, it cannot be discarded that some game-playing factor would measure competitiveness and interactivity abilities aside from deductive and inductive abilities.

Finally, we have considered individual tests which measure one factor. For measuring several factors at a time, the exercises should be given one by one and, after each guess, the subject should be given the correct answer (rewards and penalties can be used instead). This has two advantages: there is no need for the subject to understand natural language (or any language) in order to be explained the purpose of the test, and there is no need to tell which factor or purpose is to be measured in each part of the test. There is also one disadvantage, deductive problems should be posed in terms of ‘learn to solve’, and this may devirtualise them.

7 Applications

Modern AI systems are much more functional than systems from the sixties or the seventies. They solve problems in an automated way that before required human intervention. However, these complex problems are solved because a methodical solution is found by the system’s designers, not because most current systems are more intelligent than preceding ones. Fortunately, the initial aim of being more general is still represented by some subfields of AI: automated reasoning and machine learning.

Automated reasoning (more properly called Automatic Theorem Proving) is addressing more complex problems by the use of inductive techniques, while maintaining their general deductive techniques. These systems, in fact, have been used as the ‘rational core’ of many systems: knowledge-based systems, expert systems, deductive databases, ... But, remarkably, the evaluation of the growth of automated reasoning has not been established from the success of these applications but from the increasingly better results on libraries of problems, such as the TPTP library [22]. However, there is no theoretical measurement about the complexity of the problems which compose these libraries. Instead, some approximations, such as the number of clauses, use of some lemmatas, etc., have been used. Following the approach presented in this paper it would be interesting to give a value of $k - solvability$ of each of the instances of these libraries.

In a similar way, machine learning has recently taken a more experimental character and systems are evaluated wrt. sets of problems. Except from general problems (classes), where their complexity (or learnability) has been established, there is no formal framework for giving a scale for concrete instances.

In this new and beneficial interest in measurement, Bien et al. [1] have defined a ‘Machine Intelligence Quotient’ (MIQ), or, more precisely, two MIQs, from ontological and phenomenological (comparative) views. Any comparison needs a reference, and the only reference of intelligence is, for the moment, the human being. This makes the approach very anthropocentric, like the Turing Test. The ontological approach, however, is not based on computational principles but on a series of characteristics of intelligence that are defined on linguistical terms rather than computational/mathematical ones, such as long-term learning, adaptation, recognition, optimization, etc. Moreover, the evaluation is generally measured on performance on some specific problem, contrary to the claim that “it is time to begin to distinguish between general, intelligent programs and the special performance systems” [18]. Although this can be very appropriate for specific systems where functionality is clear, in general this would not allow for the comparison of intelligence skills of different systems devised for quite different goals. How to define general and absolute characteristics of intelligence computationally is more difficult and new problems present themselves, but the progress in the ‘intelligence’ of AI systems can only be measured in this way.

8 Conclusions and Future Work

Among the problems for making these measurement reliable there is the selection of a reference machine. The evaluation of abilities with instances is dangerous because it depends on constants. Since there is no apparent preference for any descriptive mechanism, we plan to adapt these notions for logic programming, because it is a paradigm that has been used both for automated

deduction and machine learning (ILP) as well as other uses (abduction, theory revision, ...), and, in our opinion, is not biased.

For the moment, the framework which has been presented allow for the measurement of different factors and clarifies the distinction between evolutionary-acquired knowledge, life-acquired-knowledge and 'liquid intelligence' (or individual adaptability). Several tests for different subfields of AI could be devised following this paradigm, and the increasing scores for solving more and more complex (k -solvable) problems may be a way to know how much intelligent AI systems are wrt. previous generations systems.

References

1. Bien, Z., Kim Y. T. and Yang, S. H., 1998, "How to Measure the Machine Intelligence Quotient (MIQ): Two Methods and Applications", *World Automation Congress (WAC)*, TSI Press, Albuquerque, NM.
2. Blum L. and Blum M., 1975, "Towards a Mathematical Theory of Inductive Inference". *Inf. and Control*, 28:125-155.
3. Bradford P. G. and Wollowski, M., 1995, "A Formalization of the Turing Test (The Turing Test as an Interactive Proof System)". *SIGART Bulletin*, 6(4), p. 10.
4. Chaitin, G. J., 1982, "Gödel's Theorem and Information". *Int. J. Theo. Phys.*, 21, 941-54.
5. Eysenck, H. J., 1979, *The Structure and Measurement of Intelligence*, Springer-Verlag.
6. Gammerman, A. and Vovk, V. (eds.), 1999, Special Issue on Kolmogorov Complexity, *The Computer Journal*, 42(4).
7. Gold, E. M., 1967, "Language Identification in the Limit", *Inform & Control*, 10, 447-474.
8. Harman, G., 1965, "The inference to the best explanation", *Philos. Review*, 74, 88-95.
9. Herken, R., 1994, *The universal Turing machine: a half-century survey*, Oxford Univ. Press, 1988, 2nd Ed., 1994.
10. Hernández-Orallo, J., 2000, "Computational Gain and Inference", *Collegium Logicum*, 4, Springer, in press.
11. Hernández-Orallo, J., 2000, "Beyond the Turing Test", to appear in *Journal of Logic, Language and Information*, to appear in Vol. 9 no. 4.
12. Hernández-Orallo, J. and Minaya-Collado, N., 1998, "A Formal Definition of Intelligence Based on an Intensional Variant of Kolmogorov Complexity" In *Proc. of the Intl. Symp. of Engin. of Intelligent Systems (EIS'98)*, ICSC Press, 146-163. 1998.
13. Johnson, W. L., 1992, "Needed: A New Test of Intelligence", *SIGART Bulletin*, 3(4), 7-9.
14. Kolmogorov, A. N., 1965, "Three Approaches to the Quantitative Definition of Information", *Problems Inform. Transmission*, 1(1):1-7.
15. Levin, L. A., 1973, "Universal search problems", *Problems Inform. Transm.*, 9, 265-6.
16. Li, M. and Vitányi, P., 1997, *An Introduction to Kolmogorov Complexity and its Applications*, 2nd Ed., Springer-Verlag.
17. Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J. Halpem, D. F., Lochlin, J. C., Perloff, R., Sternberg, R. J. and Urbina, S., 1996, "Intelligence: Knowns and Unknowns", *American Psychologist*, 51, 77-101.
18. N. J. Nilsson, Eye on the Prize. *AI Magazine*, July 1995.
19. Solomonoff, R. J., 1964, "A formal theory of inductive inference", *Inf. Control*, 7, 1-22, March, 224-254, June.
20. Solomonoff, R. J., 1978, "Complexity-based induction sytems: comparisons and convergence theorems", *IEEE Trans. Inform. Theory*, IT-24, 422-438.
21. Solomonoff, R. J., 1999, "Two Kinds of Probabilistic Induction", in the 'Special Issue on Kolmogorov Complexity', *The Computer Journal*, 42(4), 256-259.
22. Suttner, C. B. and Sutcliffe, G., 1998, "The TPTP Problem Library: CNF Release v1.2.1", *Journal of Automated Reasoning*, 21(2), 177-203.
23. Turing, A. M., 1936, "On computable numbers with an application to the Entscheidungsproblem", *Proc. London Math. Soc.*, series 2, 42, 230-65, 1936. Cor., Ibid, 43, 544-6, 1937.
24. Turing, A. M., 1950, "Computing Machinery and Intelligence", *Mind*, 59, 433-460.

ON DEFINITION OF TASK ORIENTED SYSTEM INTELLIGENCE

Michel Cotsaftis

LTME/ECE 53 rue de Grenelle 75007 Paris France

fax:33-(0)1-42-22-59-02; email:mcot@ece.fr

ABSTRACT

With development of system complexity and performances, it is important to evaluate its ability to perform tasks, especially in the case of opposing outer effects. This amounts to affect "intelligence" coefficient to the system, which basically requires to transfer usual geometric space calculations to more global and qualitative task space, the only one where this coefficient can have a meaning irrespective of system structure. The problem is discussed here by defining the useful information by its analytical expression explicit in terms of system elements. By application to the class of deformable Lagrangian systems, adapted controlled structure is constructed. Intelligence measured by minimization of a distance between demand and result mainly appears as a compromise between information ball and robustness ball reduction for fixed system complexity.

Keywords : *System complexity, Functional Asymptotic Control, Useful Information and Entropy, Intelligence, Task Space Control.*

1-INTRODUCTION

As technical systems required for real life task accomplishment are becoming very complex both in their (hardware) physical realization and in the related (software) organization of their command-control structure, an emerging question is in the possible existence of a limit in improving these systems. Supposing everything can be continuously extended on hardware side, a direct consequence on soft side is the research of a quantitative way to scale system capability, ie in short to measure their "intelligence"[1]. One should first make sure that the question has a well defined meaning as for human the definition of intelligence is multiform and depends on the emphasized "qualities" in the tests. Also, a difficulty is the domain on which this "intelligence" is applied, as there exists different kinds of human "intelligence" ranging from high abstraction to very applied domains. To avoid these problems the angle of approach will be modified and, as a system is generally designed for ac-

complishing a prescribed set of tasks, its "intelligence" compared to another system will be evaluated in terms of its "efficiency" to collect the relevant information for these tasks and to use it in its accomplishment. A companion question is system adaptation to different or even adverse working conditions, which also amounts to evaluate the size of robustness ball corresponding to the selected tasks. A difficulty however reappears with the word "selected" as concerns "who" is choosing the tasks, and this stresses the huge difference between dedicated and self-deciding system structures. In first case, "intelligence" measurement is limited to evaluation of simple faithfulness in design and organization, and to robustness to parameter change, whereas in second one, a new dimension in system evaluation capability is added, showing that the problem cannot be handled in an universal and unique framework.

Another strong restriction is coming from hardware. Example of lightweight robot arms[2] shows that for high enough power there exists a breakpoint where internal material structure generates excitation of internal deformation modes impairing initially researched performances. One may speculate that this could be cured by adequate controller design using vision system, most adapted to detect working environment and to give more flexibility to adapt to task change. As mounting vision sensor on robot arm is no longer possible with deformations, exterior more rigid fixture should be used. If environment is then correctly observed, robot arm vibrations still remain, forbidding fast enough approach to target. So including robot end effector in visual observation may appear as a natural solution, but the reality is that this is not possible as actuator frequency range is significantly smaller than typical perturbation frequency range. Active control robustification, a constant trend in control research over the last decades, becomes inefficient beyond some today crossed limit because of the unavoidable spillover from low frequency actuator range to high frequency internal system range which severely limits the performances. This internal contradiction (more controlled active power for nominally better performance

leading to secondary internal phenomena downgrading more this performance) also makes the "intelligence" assessment somewhat questionable in the present context, and bounds even more the domain where the problem has a well defined meaning. Interpretation is that usefulness of collected information from sensors is strongly system depending, including human operator, raising the problem of its adequate selection for a given system and a prescribed task.

Escaping from these difficulties is however possible by observing that this limitation comes from inability of computation-control system to reconstitute, as it classically does, actuator command from trajectory observation for its efficient control. Two different elements are implied in this statement. One is the impossibility past some level of complexity to distinguish between two close enough trajectories. Even with perfect end effector location in time and space, decomposition of this observation on base representation functions becomes unrealizable when flexion and torsion effects are mixing up in a very complicated motion. Control action becomes inefficient if one-to-one relation between control and trajectory is no longer maintained. Even if it were maintained, the power would have to be delivered, owing to speed and torque requirements, in a too high frequency range for present actuators, and this would be technically non realizable. The second element is also of fundamental nature in that there is no direct action on deformation modes from actuators, as they are receiving their power input from rigid motion mode, leading to a mismatch between internal natural power cascade and external one imposed by feedback loop with usually spillover effect impairing again system performance.

As there is inadequation between basic physics understanding and new bifurcated situation, classical point of view should be changed. With trajectory non distinguishability the base ingredient for trajectory control, ie its time dependence in usual state space representation, should be abandoned. Only trajectories as a whole have now a meaning, and global enough information is relevant. Reducing the complete non controllable system dynamics to smaller initially driving rigid ones, time dependent system trajectory is embedded into a selected class by application of fixed point theorem. The resulting control, explicitly expressed in terms of global system quantities, still gives asymptotic stability toward desired trajectory, and exhibits the interesting property to be at its level naturally organized toward task orientation. So in progressing toward higher quality performances with higher designed and more complex systems, use of better components is not sufficient and control structure has also to fit with system properties, implying mainly application of *subsidiarity principle* guaranteeing

minimization of internal information flux. This restores adjustment of system hardware structure to possible task assignment, as it gives again the system the way to have appropriate internal information exchange compatible with power flux. Resulting internal coherence thus appears as an extremely important element in the possibility of measuring system "intelligence".

To illustrate the previous concepts developed at system level, *useful* information is defined in next paragraph and task oriented control for general Lagrangian system dynamical equations is considered. Application to actuated one-link robot arm with flexion and torsion deformations carrying off-center massive object is discussed with Euler-Bernoulli approximation. When compared to usual control based on vision system which in present case cannot insure trajectory stability, "local" deformation effects are internally taken care of by proposed control. As much lower information flux circulation is implied, vision system is freed for higher level task of driving the approach to desired target, and for much more modest computing requirement. In this sense, actual system may appear as more "intelligent".

2-SYSTEM REPRESENTATION AND USEFUL INFORMATION

For global system improvement, system parts have themselves to be improved in their various components. Basically three hardware parts always exist in a system, 1)- a mechanical-physical part, 2)- a sensor-computing part, and 3)- a power-actuation part, see Fig.1. There also exists a fourth software control law part, which should enable the system to correctly perform in targetted range within its new physical conditions, manifested by the creation of a (possibly infinite) number of internal modes, thus increasing its number of degrees of freedom, and making previous classical controls inadapted. The control based on the new physical conditions theoretically exists[2] and still makes system trajectories asymptotically stable, ie it guarantees again tracking performance requirements.

Due to larger excited frequency band when mode number increases, the problem now rests upon 1)- sensing and treating this new added information, and 2)- generating the corresponding power inputs as needed for increasing system performances. The first point belongs to sensor-computing part, and is handled within existing technology covering a large frequency band with a wide set of technical solutions and corresponding to broad range of accuracy. For the second point, despite the large size domain ($[10^{-1}m, 10^1m]$) without going into more specific microsystems, there still exists a frequency gap between classical actuators low frequency domain ($[0, 30Hz]$), and high frequency domain corresponding

to "smart" material systems ($[3.10^2 Hz, 3.10^3 Hz]$). Any new information is directly usable only if it belongs to the intersection of both sensors and actuators frequency ranges. A very striking case is vision sensor giving an over-detailed amount of informations not directly useful for system control improvement. Consequently to give the system adapted capability, the problem is not in getting more information as believable from the increase of system internal degrees of freedom, but on the contrary to reduce the extra-information from state space in frequency range outside actuator's one, and in order to maintain robust asymptotic stabilization by adapted control within the uncertainty ball corresponding to the unpreciseness produced by this reduction. As shown on Fig.1, it is after collection of rough information from sensors that there should exist a reduction process to filter the only relevant information needed for reaching system targets. This leads to the definition of *useful information* determined from task orientation rather than lower level unexploitable trajectory orientation. It is based on observation that occurrence of events rests upon removal of a double uncertainty : the usual quantitative one related to occurrence probability and the qualitative one related to event utility for goal accomplishment. So events may have same probability but very different utility, and this explains why some extra informations on top of existing ones have no impact on reaching the goal. In present case, it can be verified that, calling u_j and p_j the utility and the probability of event E_j , and $I(u_j, p_j)$ its associated information called useful information, the relation

$$I(u, p_1 p_2) = I(u, p_1) + I(u, p_2) \quad (1)$$

holds for event $E_1 \cup E_2$ with same utility u . On the other hand, there is strict proportionality between utility and corresponding information, so

$$I(\lambda u, p) = \lambda I(u, p) \quad (2)$$

With eqns(1,2), there results that useful information is given by

$$I(u, p) = -ku \cdot \log p \quad (3)$$

where k is Boltzmann constant. Usual entropy calculation is thus obtained by presupposing that all events have same utility for goal accomplishment, which is certainly true in Thermodynamics where all molecules are totally interchangeable and thus indistinguishable. As a consequence it is well known that only the invariant corresponding to this equivalence class, here the energy (or the temperature), allows to separate thermodynamical systems. Similarly internal system deformations (flexion and torsion) are undistinguishable events as they are layered on invariant surfaces determined by the value of bending moment M at link's origin[3]. So using

their observation to improve system dynamical control is not possible, in the same way as observing individual molecule motion in a gas does not improve its global control. As a result, raw sensor information has to be filtered so that only useful information for desired goal is selected. This is precisely the remarkable capability of living systems to have evolved their internal structure so that this property is harmoniously embedded at each level of organisation corresponding to each level of development. In this sense they are remarkably intelligent. An important element here is that the process has been subsidised into the hardware structure in order to free the upper levels.

3-LAGRANGIAN EQUATIONS FOR DEFORMABLE SYSTEM

To proceed, advantage will be taken of the general lagrangian form of deformable system in order to exhibit directly on system equations the features discussed above concerning information reduction. First there is a cascade effect of exterior forces onto rigid dynamics feeding itself deformation modes, allowing reduction of complete initial (infinite dimensional) system to (finite dimensional) "core" rigid system, see Fig.2. Then, and as long as "natural" boundary conditions are considered for the system, only these intrinsic elements will be really needed to control system dynamics. By "natural" are meant boundary conditions constructed with the remaining terms coming from the various integrations by part needed to transform system action variation into Lagrange equations. More specifically, with Lagrangian density

$$\mathcal{L}_T = \mathcal{L}_T \left(q_j(t), \frac{dq_j(t)}{dt}, u_k(x, t), \frac{\partial u_k(x, t)}{\partial t}, \frac{\partial^m u_k(x, t)}{\partial x^m}, u_k(S_1, t), \frac{\partial u_k(S_1, t)}{\partial t}, \frac{\partial^m u_k(S_1, t)}{\partial x^m}, x, t \right) \quad (4)$$

depending on both discrete (rigid) variables $q_j(t)$ and field (deformation) variables $u_j(x, t)$ up to their p th space derivatives, as well as their values on a part (S_1) of total system boundaries ($S = S_1 \times S_2$) of the space domain $D(x)$ in the additive form

$$\mathcal{L}_T = \frac{1}{V(x)} L_R \left(q_j, \frac{dq_j}{dt}, t \right) + \mathcal{L}_D \quad (5)$$

of a rigid variable part L_R and a deformable one \mathcal{L}_D , and where the arguments in the second part are the same as in eqn(1). The variation of the action

$$\mathcal{I} = \int_{t_0}^{t_f} \int_{D(x)} \mathcal{L}_T dx dt \quad (6)$$

inside the space domain $D(x)$ and over the time interval $[t_0, t_f]$ can finally be splitted into two parts, one under the integral sign and another one expressed at the boundary (S) of $D(x)$ and at the limits of the time interval (if there are "transversality conditions"), and resulting from integrations by part. Writting that system equations are deduced from the action \mathcal{I} by a variational principle implies two elements :

- 1 - the Lagrange equations

$$\frac{\partial \int \mathcal{L}_T}{\partial q_j} - \frac{d}{dt} \frac{\partial \int \mathcal{L}_T}{\partial \dot{q}_j} = F_j + U_j$$

$$\frac{\partial \mathcal{L}_T}{\partial u} - \partial_\mu \frac{\partial \mathcal{L}_T}{\partial u_\mu} + \partial_{\mu\nu} \frac{\partial \mathcal{L}_T}{\partial u_{\mu\nu}} - \dots - \frac{d}{dt} \frac{\partial \mathcal{L}_T}{\partial \dot{u}} = 0 \quad (7)$$

are satisfied inside the space-time domain, with U_j the control acting onto the system,

- 2 - the remaining boundary terms resulting from integration by parts are equated to the work done by exterior force terms onto the system, ie.

$$\left(n_\mu \cdot \left[\frac{D\mathcal{L}_T}{Du_\mu} + \dots \right]_{S_j} + \frac{\nabla \mathcal{L}_T}{\nabla u(S_j)} \delta_{j1} \right) \cdot \delta u_{S_j} = 0$$

$$\left(n_\nu \cdot \left[\frac{\partial \mathcal{L}_T}{\partial u_{\mu\nu}} - \dots \right]_{S_j} + \frac{\nabla \mathcal{L}_T}{\nabla u_\mu(S_j)} \delta_{j1} \right) \cdot \delta u_{\mu S_j} = 0 \quad (8)$$

with

$$\nabla \mathcal{L}_T / \nabla Z = \partial \mathcal{L}_T / \partial Z(S_j) - d/dt [\partial \mathcal{L}_T / \partial \dot{Z}(S_j)]$$

$$D\mathcal{L}_T / Du_\mu = \partial \mathcal{L}_T / \partial u_\mu - \partial_\nu \partial \mathcal{L}_T / \partial u_{\mu\nu}$$

and transversality conditions if any are satisfied. Boundary conditions are called "natural" when they are constructed from these quantities, and not from different ones.

For a 1-link system, the lagrangian writes in partitioned form

$$\mathcal{L} = L_r(q_j(t), \dot{q}_j) + \int \mathcal{L}_d(q_j, \dot{q}_j, q(x, t), \dot{q}, q_\mu, q_{\mu\nu})$$

$$+ \mathcal{K}_S(q_j, \dot{q}_j, q_S, \dot{q}_S, q_{\mu S}, \dot{q}_{\mu S}) \quad (9)$$

with rigid part

$$L_r = J_a \left(\frac{d\theta}{dt} \right)^2 + J_m \left(\frac{d\theta_m}{dt} \right)^2 + K_m (\theta - \theta_m)^2 \quad (10)$$

in terms of rigid articular and actuator variables $q_1 = \theta$, $q_2 = \theta_m$, deformation part

$$\mathcal{L}_d = \rho A \left(x \frac{d\theta}{dt} + \frac{\partial u(t, x)}{\partial t} \right)^2 + \rho K^2 \left(\frac{\partial \gamma(t, x)}{\partial t} \right)^2$$

$$+ EI \left(\frac{\partial u(x, t)}{\partial x} \right)^2 + GJ \left(\frac{\partial \gamma(t, x)}{\partial x} \right)^2 \quad (11)$$

in terms of flexion and torsion variables $u(t, x)$, $\gamma(t, x)$, and interaction part

$$\mathcal{K}_S = \frac{1}{2} m X^2 + J_f \left(\frac{d\theta}{dt} + \frac{\partial^2 u(t, x)}{\partial x \partial t} \right)_{x=L}^2$$

$$+ J_t \left(\frac{\partial \gamma(t, x)}{\partial x} \right)_{x=L}^2 \quad (12)$$

$$X = (L + l_f) \frac{d\theta}{dt} + \frac{\partial u(t, x)}{\partial x} + l_f \frac{\partial^2 u(t, x)}{\partial x \partial t} + l_t \frac{\partial \gamma(t, x)}{\partial x} \Big|_{x=L}$$

at links boundaries, out of which dynamical equations and boundary conditions are easily obtained[4]. (l_f, l_t) are coordinates of tip mass m with respect to link's end, and the various other coefficients characterize the beam as usual within Euler-Bernoulli approximation. One can verify that in link and actuator equations coupled by compliance effect, are both acting the applied input torque τ and the bending moment $M_a = EI(\partial^2 u(t, 0)/\partial x^2)$, here the only term through which deformations are seen by system rigid part.

4-TASK ORIENTED CONTROL

In general, the system is assigned to perform an action, and a control is set to give the system the ability to meet the corresponding requirements. This is always expressed as satisfaction of Lyapounov theorem with adapted Lyapounov function, written in terms of system trajectory parameters in state space. In other words, control is trajectory oriented, and all sensors are used in this view. In particular, vision sensor if any will provide information on link tip motion. As seen above, this is misleading as long as observed motion belongs to an indistinguishable class. Control has to be approached in task oriented sense, and, for reaching the goal, is governed by a choice of "good" informations depending of their utility defined above. Starting from partial Hamiltonian density associated to deformable part

$$\mathcal{H}_D = \dot{q}_j \frac{\partial \mathcal{L}_D}{\partial \dot{q}_j} + \dot{u} \frac{\partial \mathcal{L}_D}{\partial \dot{u}} + \dot{u}(S) \frac{\partial \mathcal{L}_D}{\partial \dot{u}(S)} + \dot{u}_\mu(S) \frac{\partial \mathcal{L}_D}{\partial \dot{u}_\mu(S)} - \mathcal{L}_D \quad (13)$$

one will consider system Lyapounov function

$$\mathcal{V} = \int \mathcal{H}_D + \sum_j \left(K_{Pj} \frac{q_j^2}{2} + \Gamma_{Vj} \frac{\dot{q}_j^2}{2} \right) \quad (14)$$

with positive parameter gains K_{Pj}, Γ_{Vj} . Its time derivative along system trajectories is

$$\frac{d\mathcal{V}}{dt} = \sum_j \dot{q}_j \left[U_j + F_j - \left(\frac{\partial L_R}{\partial q_j} - \frac{d}{dt} \frac{\partial L_R}{\partial \dot{q}_j} \right) \right]$$

$$+ K_{Pj} q_j + \Gamma_{Vj} \ddot{q}_j \quad (15)$$

Substituting for d^2q_j/dt^2 from explicit Lagrange equations(7) and eliminating all other second order time derivatives, one will get an "inertia" term F_{aj} which, on physical grounds, is equal to forces other than exterior forces F_j acting onto system of discrete variables q_j , and coming from the (back) effect of the field variables $u(x, t)$ onto discrete variables $q_j(t)$. As \mathcal{V} is positive definite for large enough definite positive gains (K_{Pj}, Γ_{Vj}) , its derivative can be made definite negative by taking control U_j so that the term between brackets is equal to $-(K_V \dot{q})_j$, where matrix K_V is definite positive. The resulting form of the control (supposing there is no exterior force)

$$U_j = -K_{Pj}q_j - K_V\dot{q}_j + K_D(q_j, \dot{q}_j, \dots) + K_{Fj}F_{aj} \quad (16)$$

and generalizes usual PD-control to full nonlinear case. In fact, it fits more generally the expression of dynamical system control

$$U = U_{comp} + \overline{U}_{PDF} + \Delta U \quad (17)$$

when writing the tracking condition for desired trajectory $q_j(t) = q_{ja}(t)$ and splitting the various control components, with

$$U_{PDF} = \overline{U}_{PD} + K_F \begin{bmatrix} 1 \\ 0 \end{bmatrix} F_a \quad (18)$$

Moreover, from argument above, the control law in eqn (16) gives both asymptotic tracking of desired trajectory for discrete variables and asymptotic stability for field variables as well as their first order time derivatives.

From eqn(15), equating the sum between brackets in its right hand side to $-(K_V \dot{q})_j$ amounts to take a controller of PDA type[5]. However, it should be observed that the resulting invariant subset of $d\mathcal{V}/dt$ is the same as when $\Gamma_j = 0$. So the same convergence property of the solutions is expectable for any value of Γ_j . The reason of introducing the new kinetic term with $\Gamma_j \neq 0$ is in the role of the direct acceleration term, or of the new resulting term F_{aj} after substitution, which is mainly to change the relative values of inertia-damping-stiffness system parameters with respect to field modes, as already observed and used for classical force control.

But after substitution from Lagrange equations this term is an integral of a complicated function of field variables and their space derivatives over the domain $D(x)$. So there is no advantage to use it in this form which requires local knowledge of field variables inside the domain, unless Lagrangian structure is such that this integral transforms into explicit well identified and sometimes directly measurable surface quantities. A very simple case occur when the Lagrangian \mathcal{L}_D is such that formally

$$\frac{\partial \mathcal{L}_D}{\partial q_j} - \frac{d}{dt} \frac{\partial \mathcal{L}_D}{\partial \dot{q}_j} = \frac{\partial \mathcal{L}_D}{\partial u} - \frac{d}{dt} \frac{\partial \mathcal{L}_D}{\partial \dot{u}} \quad (19)$$

Then from Lagrange eqns(7) there results

$$-\left(\frac{\partial L_R}{\partial q_j} - \frac{d}{dt} \frac{\partial L_R}{\partial \dot{q}_j}\right) = n_\mu \cdot \left[\frac{\partial \mathcal{L}_T}{\partial u_\mu} - \partial_\nu \frac{\partial \mathcal{L}_T}{\partial u_{\mu\nu}} + \dots\right]_{S_2}$$

The "inertia" force term F_{a2} is just equal to the boundary term in the first bracket of eqn(8) when $\Gamma_j = 0$, and is expressible in terms of this quantity, and of rigid variables q_j and their first time derivatives when $\Gamma_j > 0$. This global expression contains all needed information to control the local action of (infinite dimensional) deformation effects, usually approached by decomposing this source term onto all projection space and cutting at a finite mode number with spillover consequences[6,7].

Much more than local control, more global task oriented control will also be independent of (too) detailed information on link deformations. Typical task is to reach a preassigned target under specific circumstances. Returning to eqn(3), this amounts to minimize the total entropy production associated to any motion in the class of acceptable trajectories fixed by the local control defined in previous paragraph, so its expression depends in general of all trajectory parameters. To this end, the utility u will be taken as the gradient of a convenient positive definite quantity such as a Lyapounov function to define a steepest path and more importantly, to eliminate before data processing irrelevant task information, saving enormous amount of time and data space. So with (p) the set of all observed parameters one gets

$$u = \frac{\partial \mathcal{V}}{\partial p} \quad (20)$$

and in eqn(3) only will remain terms for which this expression is above a minimum threshold value corresponding to system sensitivity. So all collected information from sensors is filtered in terms of its utility for the prescribed task. This explicit result is independent of the dedicated or selfdeciding character of the system. With eqn(14) for instance, the only dependence of \mathcal{V} on trajectory parameters is through bending moment M , so when taking the gradient with all sensor information, there only remains a term $\partial \mathcal{V}/\partial M$, and more detailed trajectory information does not appear. So adapted control splits finally into a local one expressed in terms of global (relative) invariants M , and a nonlocal one depending on utility of these quantities for reaching final target. Though trajectory oriented the first one directly links to the task oriented second one and respects the very nature of internal information provided by system structure. In this respect, system intelligence is easily measured by information flux from eqn(3) and by associated robustness ball of the applied control corresponding to a distance between demand and result.

5-CONCLUSION

Analysis of system structure shows that evaluation of its intelligence is only meaningful in task space. This requires the satisfaction of internal coherence conditions manifested by system ability to extract from its sensors the relevant information for these tasks. The problem is studied here by defining the useful information which precisely allow to pass from initial geometrical space to task space irrelevant of the way the system is designed and organized. Application is made for Lagrangian systems representing deformable bodies, for which equations analysis shows that even if at first sight system nature is drastically changing with increase of state space dimension to infinity, internal system organization also changes in such a way that its local control still remains fundamentally finite dimensional. Observation of new deformation modes is not only useless, but also damaging in that it leads to control form interfering with natural internal feedback regulating power exchange between displacement and deformation. Sensors providing too detailed information are not adapted as it has to be filtered for reconstitution of needed more global one. More efficient way is to use local control based on natural system invariants, directly linkable to more global task oriented control based on useful information (rather than filtered one) expressed in terms of utility factors constructed as the gradient of Lyapounov with respect to trajectory parameters. When they aggregate into trajectory invariants, only their derivatives finally

appear, justifying again the choice of previous local control form. Moreover, the association of the two level form presented here respects natural system organization and minimizes information transfer between the two levels. System intelligence is directly measured by task adaptation expressed here as both circulating information flux and robustness ball corresponding to local controller for a given distance between demand and result.

References

- [1] J. Khalfa, Ed. : *What is intelligence*, Cambridge University Press, Cambridge, Mass., 1994
- [2] W.J. Book, T.E. Alberts, G.G. Hastings : "Design Strategies for High-speed Lightweight Robots", *Computers in Mechanical Engineering*, p.26,1986.
- [3] M. Cotsaftis : *Comportement et Contrôle des Systèmes Complexes*, Paris, Diderot Multimédia, 1997.
- [4] M. Cotsaftis : "Global Control of Flexural and Torsional Deformations of One-link Mechanical Systems", *Kybernetika*, Vol.33(1),1997,pp.75-86.
- [5] P.T. Kotnik, S. Yurkovitch, U. Ozguner : "Acceleration Feedback Control for a Flexible Manipulator Arm", *Proc. 1988 IEEE Intern. Conf. on Robotics and Automation*, Philadelphia, Penn.,Vol.1,1988,p.322
- [6] M.J. Balas : "Modal Control of Certain Flexible Dynamic Systems", *SIAM J. Control*, Vol.16,1978,p.450.
- [7] Y. Sakawa : "Feedback Control of Second Order Evolution Equation with Unbounded Observation", *Int. J. of Control*, Vol.41(3),1985,p.713.

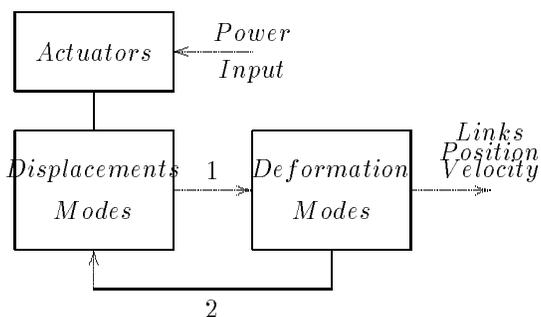


Fig.2 : Complex System Structure of Deformable Mechanical System

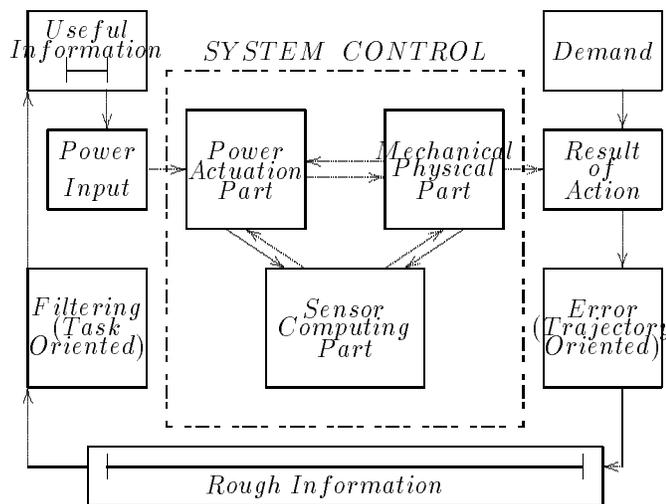


Fig.1 : System Structure with Main Component Parts and Information Filtering for Task Orientation
 Problem : Minimize distance(demand,result)
 $=f(\text{system parameters})$

Minds, MIPS and Structural Feedback

Ricardo Sanz and Ignacio López

Universidad Politécnica de Madrid

{sanz, ilopez}@etsii.upm.es

Abstract

This paper tries to stress the need of having a clear understanding of the concept of intelligence before we can progress in the formulation of a measure for it. At the end it suggests a view of intelligence as *structural feedback in model-based control systems*.

Keywords: *Intelligence, performance, behavior, mental models, structural feedback.*

1 INTRODUCTION

This paper tries to suggest the practical impossibility of finding a *single* and *useful*¹ measure of general intelligence for all types of artificial systems performances unless we get some previous result in the form of a sound theory of intelligence.

As was stated at the workshop website, its goal is to discuss three challenges pertaining to intelligent system performance:

- how to measure performance;
- how to evaluate intelligence and
- how to put performance and intelligence into correspondence.

We will try to address the three points in order (see sections 4,5 and 6), but first we want to make a first comment. When talking about intelligence a problem appears, and it is that "intelligence" is a moving target. Some centuries ago "a person able to read" implied "a person very intelligent". Now we don't consider this ability as a symptom of intelligence in a person of our environment. But if we talk about an animal, for some example a dog, "able to read" is still considered a good manifestation of intelligence.

So, what is that stuff that appears or disappears as you point at different entities? Can intelligence be in the eye of the beholder? We think that the term is used in two quite different ways: a) As a comparison between two entities that can be both explicit or one implicit (a normal dog) and b) As an absolute measure of some core capability.

While we can mostly agree with Alex Meystel conception of intelligence as a *core concept underlying minds*, perhaps all we

¹From an engineering point of view, *i.e.* to build/analyze artificial systems.

are falling in the easy way of thinking mentioned by Bateson [3, page 82] of *using words that appear more concrete than they are*².

Before entering into main matter, let's start with a brief discussion about the adequacy of ascribing mental properties like *intelligence* to machines.

2 WHAT IS INTELLIGENCE?

It is common to address intelligence as a property inherent to something we call mind. The use of both terms, intelligence and mind, is not that clear. In fact, each one of us appears to have his own notion of intelligence speaking in terms of everyday life. Although deep thought and study about the topic can clarify partial notions of intelligence, there is still no global perspective.

We want the following question to emerge: *does intelligence really exist?* After what has been said and having in mind our constant references to the concept, it really seems ridiculous to question it. But we would like to point out the fact that *intelligence* could well be one word hiding what can be considered a too fuzzy concept³. By this we mean that the word does not have a fixed reference to something that can be pointed out, such as a dog or a table (it lacks a true referent). It is in some sense a concept similar to a notion of a mathematical space, *i.e.*: everything which matches certain restrictions is part of *intelligence*. The space of things that think.

The concept has lost in this way the apparent rigidity; the question, although, may be, in a more precise way: *what are the restrictions a feature has to match to form part of intelligence?* And at this point the answers diverge because the number of possibilities is close to infinity.⁴ It would be an error to put the question like this. Perhaps it would be better to approach the topic in another way: *what is behind everything we seem to consider intelligent?* Searching this instead of a particular set of characteristics would eventually lead to a rule with which the judgement of the existence of intelligence would be possible.

In any case, once it is clear if something is intelligent or not, it would be tempting to determine *how intelligent*, that is, *how much intelligence it has*. This question is too particular to be

²Bateson says about these words that they are too short and this shortness conveys an erroneous ascription of concreteness.

³A *linguistic variable* in its most pure sense: *i.e.* created by language.

⁴This is to be thought in a sense of *too broad for understanding*.

answered. The individual intelligent characteristics which constitute the *intelligent set of features* one self possesses are each specialised, and in this way not comparable.⁵ In this way, given a set of intelligent characteristics, the only judgement that has any sense needs to be put in terms of targets and adequation to those targets: performance.

Returning to the rule which would enable discrimination between intelligent and not intelligent, it should not be focused on common aspects of features we usually consider intelligent, but on requirements which make them possible. For example parallel calculation, memory, etc. Having this in mind, the decision to consider something intelligent or not comes from the process of analysis of the underlying capability, i.e.: learning what can be expected from a being with such capability (eg. memory) when in a particular environment and with a more or less elaborate set of targets. Apparently we end again with a certain notion of performance.

The last point we would like to focus on comes from looking at the problem from a different angle. What if *intelligence* were a concept only suitable -clear enough- for human minds? That is, we call *something* intelligence, but it does not seem to have a bounded notion behind. So, supposing it is a collection of features we have grouped together, and not considering the fact that we could have done so in other ways, what makes us think that intelligence *is* something (a table, a bus)? In other words, what makes us think an alien would have a notion parallel to our *intelligence* as he would if he came to Earth and saw a table or a boat?

3 HUMAN (SPECIES) CHAUVINISM

Let's see what philosophers think about mental properties of machines. An example is what Crockett [5, p.193] says about the use of human-like phrases to refer to machine thinking:

Our anthropomorphizing proclivity is to reify those abstractions and suppose that the computer program possesses something approximating the range of properties that we associate with similar abstractions in human minds.

Even more amazing is his continuation:

This is harmless so long as we remember that such characterizations can lead to considerable philosophic misunderstanding.

What amazes me more in this text is that people like Crockett strongly believe that *we know* what are the "abstractions in human minds" but only *suppose* what the computer program possesses. In our experience we know -most of the time- what are the abstractions -the representations- in mechanical minds but only suppose what are those abstractions in biological minds.

⁵It would be like comparing -adding, subtracting, etc.- apples and dogs: impossible.

It is these days is when we are starting to get some direct insight into the inner working of human minds by means of PET (positron emission tomography) or fMR (functional magnetic resonance [4]). As an example, fMR has confirmed what many had long suspected -that men and women think differently. Yale Medical School investigators did compare the brain operation of men and women while reading, discovering different activation patterns in their brains while performing the reading task.

Another example of the difficulties in matching human mental concepts with machine mental concepts can be found in [2]:

Indeed, if mechanical devices can distinguish wavelengths of light without having sensations, then why do I experience any sensation at all?.

Most people tend to think that the human *sensation* is something more than the mere recognition of an input signal. Recognition at the simple level of signal capture, representation and triggering of activity. "Sensation" is nothing more than the triggering of activity due to an input signal. The immediate implementation in a computer is as an interrupt handler. The only difference is the high level of concurrence in biological computers that let them be truly concurrent in responses to sensations. There are also human sensations that are so strong that they disable further sensations. This is, exactly, the type of behavior found when a computer interrupt handler disables further interruptions.

Computers provide minds for physical systems, and it is time to clarify the true meaning of *mental concepts*.

4 PERFORMANCE AND MIPS IN BRAINS

A visible feature of biological intelligence is *performance* as Jim Albus pointed in his definition of intelligence. This is related to how we use the term for humans (remember the title of the book by Sternberg and Wagner, *Practical Intelligence: Nature and Origins of Competence in the Everyday World*).

In our search for metrics for intelligence, we are exactly in the same situation as computer consumers and manufacturers were some decades ago in relation with client-requested performance measures. As they both discovered, the old-basic measure of performance (MIPS: Million Instruction per Second) was useless to compare different architectures (*e.g.* CISC vs. RISC) or applications (*e.g.* data-bases vs. finite-element analysis). The only *useful* possibility they found was the evaluation of the performance in specific tasks, and hence this was the origin of benchmarking. Unfortunately benchmarks are not single measures, and attempts to build weighted benchmarks only changed the focus of the benchmark but not the final usability of them (they are always measures of niches of functionality).

Task-independent measures, like MIPS or *bits/second* or *entropy*, are too raw to be useful for most engineering purposes because they are so far from the desired performance specifica-

tion that we lack a theory that can map one into another⁶. For example, suppose that we want a distillation column controller intelligent enough to minimize recirculation (a desired performance). Who can decide, based on a MIPS-like measure, if a fuzzy controller A can fulfill the task, or if model-based predictive controller B is better than A?

This theory that maps a *MIPS-like measure to performance specified in useful terms* is what we are seeking in our research on intelligent systems, because it is –in fact– *The True Theory of Intelligence*. The theory will not only let us evaluate alternative designs, it will be a true explanatory discourse that will reduce intelligence to simpler, well grounded, terms.

To follow Bateson suggestion of marking concepts that are not concrete enough and require further thinking, we can use the term *i-stuff* to refer to the substance measured by True Intelligence Metrics. George Saridis probably will equate *i-stuff* to negentropy and Jim Albus to performance. We will make a suggestion at the end of the paper.

5 INTELLIGENCE AND BODILY CAPABILITIES

In relation with what can we measure, we agree with Chris Landauer in the fact that "Success is not by itself the right criterion" because we have to split success into two contributions: mind and body (and bodily intelligence is not what we are talking about). As an example consider two implementations of a future Mars rover whose main mission is going from point A to point B, one kilometer away, taking a sample of the ground each 50 meters:

Implementation H: 200 Ton. Caterpillar structure based on a combination of bulldozer, power shovel and truck. Control of sample taking based on mechanical coupling of power shovel to caterpillar (50 meters = sample). It lacks directional control because it is not necessary (it will advance straight *bulldozing* any obstacle.)

Implementation T: 50 Kilograms. 10 Watt solar power panel. Microrobotic arm.

Who will attain success? If both are successful, who is more intelligent? Is performance a manifestation of intelligence? The two first questions are rhetoric. The answer for the last one is "not always".

There are some attempts to extend fundamental physical theory to include information at the same level as mass and energy. In some sense we can analyze biological behavior as an exchange of mass (feeding in / excreting out), energy (chemical in / thermal & mechanical out) or information (sensing in / speech out). We can attach these interchanges to human subsystems, and information will become associated to the mental system. This division is, however, not very strict, because information is supported by means of mass or energy, and some energy inputs are managed as mass inputs (specially in animals).

⁶This is, in fact, the third point mentioned in the introduction.

6 CONCLUSIONS

Our analysis of the Mars rover story is that if the T implementation is successful everybody will agree that it is more intelligent than the H implementation. Even if both attain success. TO achieve this result the T implementation needs some mental content and some algorithms to exploit this mental content.

As we did say before we will propose a different interpretation of *i-stuff*: it is focused on *mental models*. Following this idea, an intelligent being is a being that has models of his world in his mind and achieves intelligent behavior using its models for action. Intelligence is, from this perspective, a two sided concept: model-based mental content (static view of intelligence) and model-based generation of behavior (dynamic view of intelligence).

Can the *i-stuff* be that collection models? Not so. Because all we know some knowledgeable people that are plain stupid.

What we consider the true core of intelligence is -plainly-feedback. When feedback for action is done trough good models of the world it achieves incredible performance levels. When feedback is used to tune parameter models it make systems adapt to changing circumstances in the world. When feedback is used to modify models of the world this is a pure learning process. When feedback is used to structurally modify the algorithms exploiting the models we are talking of creativity⁷. *Structural feedback* is perhaps the highest manifestation of intelligence; when a system is able to create new control policies that will enhance its effectiveness.

Perhaps this proposal only muddles more the discussion because *model* is even shorter than *intelligence* and it seems even more concrete; but we think that it is relatively easier to devise metrics for model quality.

But even if we can measure quality of models and model evolution algorithms, we are still halfway to the metric of intelligent behavior, because we still lack a quality measure of the use of the model to generate the behavior (*i.e.* a metric of the architecture). Performance-based metrics, as suggested by Jim Albus definition of intelligence, will fit this niche but still they will be domain-dependent.

We strongly believe that, in the future, all these theories of intelligence will consolidate in a Great Unification Theory (and this *structural feedback* seems to us a good promising starting point), that will let engineers build artificial intelligences with the plasticity enough to adapt or tune to specific needs. Being this the case, in our opinion the core foundation of it will be raw information processing with capability to autoorganize in the form of models of the world and model exploitators generating behavior. The theory of intelligence can be viewed as a theory of action, a theory of representation or both.

⁷Adaptation, learning, evolution, creativity, are facets -i.e. perceptions from an external entity- of a system changing in response to interactions with the world.

References

- [1] Christian Balkenius. *Natural Intelligence in Artificial Creatures*. PhD thesis, Lund University, Lund, Sweden, 1995.
- [2] Stephen M. Barr. A mystery wrapped in an enigma. *First Things: A Monthly Journal of Religion and Public Life*, (77), November 1997. A comment on *The Conscious Mind: In Search of a Fundamental Theory*. By David J. Chalmers. Oxford University Press.
- [3] Gregory Bateson. *Steps to an Ecology of Mind*. The University of Chicago Press, 1972.
- [4] Neil R. Carlson. *Physiology of Behavior*. Allyn & Bacon, sixth edition, 1998.
- [5] Larry J. Crockett. *The Turing Test and the Frame Problem. AI's Mistaken Understanding of Intelligence*. Ablex Publishing Corporation, Norwood, N.J., 1994.
- [6] Kenneth M. Ford, Clark Glymour, and Patrick J. Hayes, editors. *Android Epistemology*. MIT Press, Cambridge, MA, 1995.
- [7] Stanley P. Franklin. *Artificial Minds*. MIT Press, Cambridge, MA, 1995.
- [8] Markus P.J. Fromherz, Vijay A. Saraswat, and Daniel G. Bobrow. Model-based computing: Developing flexible machine control software. *Artificial Intelligence*, 114(1-2):157–202, October 1999.
- [9] Donald Gillies. *Artificial Intelligence and Scientific Method*. Oxford University Press, Oxford-New York, 1996.
- [10] John Haugeland, editor. *Mind Design II*. MIT Press, Cambridge, MA, 1997.
- [11] Tariq Samad. Complexity management: Multidisciplinary perspectives on automation and control. Technical Report CON-R98-001, Honeywell Technology Center, Minneapolis, MI, January 1998.
- [12] Ricardo Sanz, Fernando Matía, and Santos Galán. Fridges, elephants and the meaning of autonomys and intelligence. In *Proceedings of IEEE ISIC'2000*, Patras, Greece, 2000.
- [13] Luc Steels and Rodney Brooks. *The Artificial Life Route to Artificial Intelligence*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1995.
- [14] Kurt VanLehn, editor. *Architectures for Intelligence*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991. The 22nd Carnegie Mellon Symposium on Cognition.

Fast Frugal and Accurate – the Mark of Intelligence: Towards Model-Based Design, Implementation, and Evaluation of Real-Time Systems

Bernard P. Zeigler and Hessam S. Sarjoughian
Arizona Center for Integrative Modeling & Simulation
Electrical & Computer Engineering Department
University of Arizona, Tucson, AZ 85721-0104, USA
Email: {zeigler|hessam}@ece.arizona.edu
URL: www.acims.arizona.edu

ABSTRACT

Engineered systems, whether called intelligent or not, principally must rely on models to achieve their goals even in the simplest situations. Therefore, a system's intelligence is a consequence of the collective intelligence embodied in its models. In this paper, we describe intelligence measurement grounded in the general concepts of discrete event, model-based system design methodology. We discuss the basic elements of the approach in view of their role in intelligence measurement. Computational resources in both processing and communication forms are constraints on intelligence, but they are not determinant. The architecture which configures these resources plays a major role in the intelligence achieved. Further the architecture must support fast and frugal heuristics tuned to the environments in which the system is to operate. Real time processing architectures built on discrete event modeling and simulation principles are most suited to support "fast frugal and accurate" intelligence. Such architectures must be designed with a software engineering methodology that explicitly supports a system's control of its own computational resources and includes hooks for measuring its intelligence in terms of the speed, frugality and accuracy of its responses.

1 INTRODUCTION

Unless we are talking about the affluent life known to many of us in the recent past, the real world is a threatening environment where knowledge is limited, computational resources are bounded, and there is no time for sophisticated reasoning. Unfortunately, traditional models in cognitive science, economics, and animal behavior have used theoretical frameworks that endow rational agents with full information of their environments, unlimited powers of reasoning and endless time to make decisions. Tacitly accepting this paradigm – as seems the prevalent assumption – does not provide a promising basis for measuring intelligence, the theme of this conference.¹ Indeed, to measure intelligence requires first an understanding of the essence of intelligence as a problem solving mechanism dedicated to the life and death survival of organisms in the real world. Evidence and theory from disparate sources have been accumulating that offer alternatives to the traditional paradigm.

¹ NIST Workshop on Performance Metrics for Intelligent systems.

An important crystallization of the new thinking is the "fast frugal and accurate" (FFA) perspective on real world intelligence promoted by Todd and Gigerenzer [1]. FFA heuristics are simple rules demanding realistic mental resources that enable both living organisms and artificial systems to make smart choices quickly with a minimum of information. They are accurate because they exploit the way that information is structured in the particular environments in which they operate. Todd and Gigerenzer show how simple building blocks that control attention to informative cues, terminate search processing, and make final decisions can be put together to form classes of heuristics that have been shown in many studies to perform at least as well as more complex information-hungry algorithms. Moreover, such FFA heuristics are more robust than others when generalizing to new data since they require fewer parameters to identify.

It is important to note that FFAs are a different breed of heuristics. They are not optimization algorithms that have been modified to run under computational resource constraints, e.g., tree searches that are cut short when time or memory run out. Typical FFA schemes enable ignorance-based and one-reason decision making for choice, elimination models for categorization, and satisfying heuristics for sequential search. Leaving a full discussion of the differences to [1], the critical distinction is that FFA's are structured from the start to exploit certain restrictive assumptions, such as skewed frequency distributions, about their input data. They work well because these assumptions often happen to hold for data from the real world. Thus FFAs are not generic inference engines operating on specialized knowledge bases (the paradigm of expert systems) nor other generalized processing structures (e.g., [2]) operating under limited time and memory constraints. An organism's FFAs are essentially *models* of the real environment in which it has found its niche and to which it has (been) adapted.

New kinds of models for biological neurons provide possible mechanisms for implementing intelligence that is characterized by fast, frugal and accurate heuristics. Work by Gautrais and Thorpe [3] has yielded a strong argument for "one spike per neuron" processing in biological brains. "One-spike-per-neuron" refers to information transmission from

neuron to neuron by single pulses (spikes) rather than pulse trains or firing frequencies. A face recognition multi-layered neural architecture based on the one-spike, discrete event principles has been demonstrated to better conform to the known time response constraints of human processing and also to execute computationally much faster than a comparable conventional artificial neural net [4]². The distinguishing feature of the one-spike neural architecture is that it relies on a temporal, rather than a firing rate, code for propagating information through neural processing layers. This means that an interneuron fires as soon as it has accumulated sufficient input "evidence" and therefore the elapsed time to its first output spike codes the strength of this evidence. In contrast to conventional synchronously timed nets, in fast neural architectures single spike information pulses are able to traverse a multi-layered hierarchy asynchronously and as fast as the evidential support allows. Thorpe's research team has also shown that "act-as-soon-as-evidence-permits" behavior can be implemented by "order-of-arrival" neurons which have plausible real world implementations. Such processing is invariant with respect to input intensity because response latencies are uniformly affected by such changes. Moreover, coding which exploits firing order of neurons is much more efficient than a firing-rate code, which is based on neuron counts [3,4].

Countering the evidence that intelligence is essentially fast, frugal and accurate is Hans Moravec's prediction that by 2050 robot "brains" based on computers that execute 100 trillion instructions per second (IPS) will start rivaling human intelligence [5]. Underlying this argument is that there is an equivalence between numbers of neurons in biological brains and IPS in artificial computers. It takes so many billions of neurons to create an intelligent human and likewise so many trillions of IPS to implement an intelligent robot. In strong form this equivalence implies that pure brute force can produce intelligence and the structures, neural or artificial, underlying fast and frugal processing are of little significance.

2 MODEL-BASED INTELLIGENCE AND MEASUREMENT

In this section, we discuss intelligent systems from three perspectives: knowledge representation, execution, and measurement. Specifically, this paper makes the case that³

- computational resources in both processing and communication forms are constraints on intelligence, but they are not determinant
- the architecture which configures these resources plays a major role in the intelligence achieved
- the architecture must support fast and frugal heuristics tuned to the environments in which the system is to operate
- real time processing architectures built on discrete event modeling and simulation principles are most suited to support FFA intelligence
- such architectures must be designed with a software engineering methodology that explicitly supports a system's control of its own computational resources and includes hooks for measuring its intelligence based on FFA standards.

2.1 *Computational resources in both processing and communication forms are constraints on intelligence, but they are not determinant*

Morevac's claim that artificial intelligence will arise once the processing power is there to support it can be the starting point for a serious investigation to understand its merits. On the one hand, we need yardsticks of intelligence and on the other, yardsticks of computational resources (presuming that raw IPS is not very discerning). We might have a diagram as shown in Figure 1.

Let's assume for a moment that we have the framework in the form of a diagram as above, what can we do with it? We can ask

- For a given level of resources, how smart can a system be? This would prevent us from trying to build systems that are infeasible with the resources at hand.
- For a given intelligence level, how much resources are needed? This would help provide cost estimates for given intelligence requirements.
- How well does a system utilize its resources? Where does its intelligence stand relative to the best achievable in its resource league? Where does its level or resources stand relative to the best in its intelligence class?

² The face recognition layered net was executed by a discrete event simulator and took between 1 and 6 seconds to recognize a face on a Pentium PC vs. several minutes for a conventional net on a SGI Indigo. Recognition performance in both cases was very high. The authors employed a training procedure which, while effective, is not plausible as an in-situ learning mechanism.

³ We are not claiming that these are the only elements responsible for intelligent behavior and by implication there are other means for intelligence measurement.

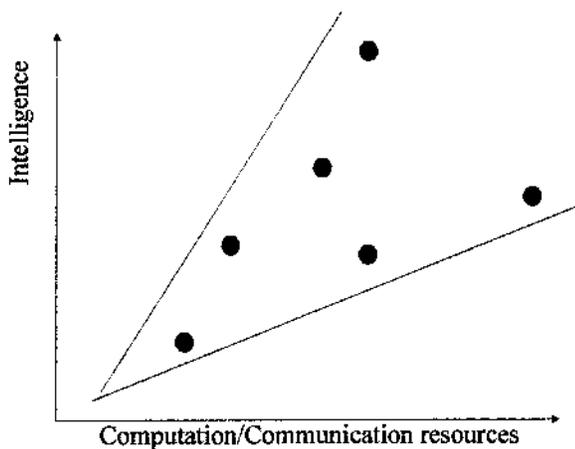


Figure 1: Intelligence measurement in terms of required resources

However, the yardsticks for resources and intelligence are not likely to be single dimensional linear orders but more likely to be multidimensional, partial orders. Even more to measure FFA intelligence which is environment-dependent, we may need to condition measurement with respect to problem classes asking which kinds of problems are performable on which kinds of architectures.

2.2 *The architecture which configures these resources plays a major role in the intelligence achieved*

This is a truism when applied to implementation of standard functionality – certain designs are better than others in implementing the same input/output behavior. However, in the absence of a well-defined characterization of intelligence in terms of input/output behavior, the focus has so far been on achieving intelligent behavior by whatever means possible, not paying much attention to the critical nature of the architectures that can support it. The results of Thorpe mentioned above, however, suggest that FFA intelligence is only achieved with “single-spike” neuron architectures and would be infeasible if the same neurons were employed in the manner assumed in conventional connectionist approaches.

2.3 *The architecture must support fast and frugal heuristics tuned to the environments in which the system is to operate*

Generalizing the idea that FFA heuristics embody models of the environment, the ability to work with models of the environment, one’s self and others may be taken as key component of intelligence. Model-based design was formally introduced around 1980s as the basis to enable systems to reason about their own behavior in normal as well as abnormal situations. Over the years, many architectures have been

proposed and implemented most of which typically suitable for narrow well-defined domains. However, a generic architecture based on simulation modeling concepts was proposed by [6]. Briefly stated, generic model-based design provides a generally applicable architecture in which simulation and other engines execute models that embody what the system employs about its environment – both external and internal

2.4 *Real time processing architectures built on discrete event modeling and simulation principles are most suited to support FFA intelligence*

Discrete event models can be distinguished along at least two dimensions from traditional dynamic system models – how they treat passage of time (stepped vs. event-driven) and how they treat coordination of component elements (synchronous vs. asynchronous). Recent event-based approaches enable more realistic representation of loosely coordinated semi-autonomous processes, while traditional models such as differential equations and cellular automata tend to impose strict global coordination on such components. Discrete event concepts are also the basis for advanced distributed simulation environments, such as the High Level Architecture (HLA) of the Department of Defense, that employ multiple computers exchanging data and synchronization signals through message passing [7]. Event-based simulation is inherently efficient since it concentrates processing attention on events – significant changes in states that are relatively rare in space and time – rather than continually processing every component at every time step.

The DEVS (Discrete Event Systems Specification) formalism [8] provides a way of expressing discrete event models and a basis for an open distributed simulation environment [9]. DEVS is universal for discrete event dynamic systems and is capable of representing a wide class of other dynamic systems. Universality for discrete event systems is defined as the ability to represent the behavior of any discrete event model where “represent” and “behavior” are appropriately defined. Concerning other dynamic system classes, DEVS can exactly simulate discrete time systems such as cellular automata and approximate, as closely as desired, differential equation systems. This theory is presented in [8, 10]. It also supports hierarchical modular construction and composition methodology [11]. This bottom-up methodology keeps incremental complexity bounded and permits stage-wise verification since each coupled model “build” can be independently tested.

An abstraction is a formalism that attempts to capture the essence of a complex phenomenon relative to a set of behaviors of interest to a modeler. A discrete event abstraction represents dynamic systems through two basic elements: discretely occurring *events* and the *time intervals* that separate them (Figure 2). It is the information carried in events and their temporal separations that DEVS employs to approximate

arbitrary systems. In the quantized systems [8], events are boundary crossings and the details of the trajectories from one crossing to another are glossed over with only the time between crossings preserved.

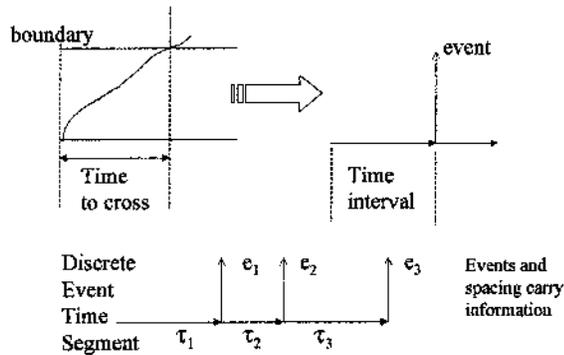


Figure 2: Discrete event representation of continuous trajectories

Recent results on discrete event neurons⁴ show that, using a race analogy, a net of simple discrete event neurons can find the shortest path in a graph in the shortest time possible. Here is an instance where fast and frugal is provably optimal! In contrast, finding the longest path (or a long path) is much more difficult and requires much more sophisticated neurons. It seems uncanny – indeed, counterintuitively so – that minimizing performance measures such as distance, time, or cost requires simple apparatus and can be done with full accuracy and without backtracking. As with FFA heuristics, the mystery dissolves when one recognizes that the discrete event neural nets exploit the underlying nature of reality in which pulses compete in parallel, and where fast competitors come first and lock out their slower counterparts from further progress. In the real world, fast response is paramount⁵ and so minimizing time (or other measures mapped into it) is critically important to survival. So brains may have been evolved to solve survival-critical problems with frugal means (simple neurons) that embody race analogies. Finally we note that discrete event neurons and one-spike-per-neuron architectures are necessary to embody the race analogy – other models do-not work.

2.5 Such architectures must be designed with a software engineering methodology that explicitly supports a system's control of its own computational resources and includes hooks for

measuring its intelligence based on FFA standards

Based on a wealth of basic research in a variety of disciplines, model-based design offers not only well-defined principles to design intelligent systems, but also can provide the means to assess a system from its inception to realization, operation, and eventual retirement. For example, we can assess a system's correctness, performance, maintenance, and cost, all of which are reflections of a system's degree of intelligence. We may also rank a system degree of intelligence in terms of, for example, intelligence of embodied models and how intelligently physical resources (computational and communication resources) are used.

Model-based design suggests several ways to rank intelligent systems based on their use of models:

- Distributed heterogeneous model-based architectures rank higher than monolithic ones.
- Systems that employ models that are at a resolution level compatible with the resources available to interpret them rank above those that don't.
 - Model sets that include self-representation rank above those that don't
 - Model sets that include representation of self and others rank above those that include only self-representation.
- Other rankings may be based on
 - Model abilities to handle both non-linguistic and linguistic queries
 - System ability to maintain coherence in the model base
 - System ability to inform meta-level models by questioning lower level models
 - Recursive depth of the "models-of" relation.

Due to increasing complexity and size/scale of systems (e.g., distributed agent-oriented systems), it is becoming imperative to follow well-defined software development processes (e.g., waterfall, spiral, iterative, and/or incremental process [12]). A typical software development process is composed of conceptualization, analysis, design, implementation, and testing, and operation [13]. Indeed, the development of many contemporary distributed, heterogeneous systems must increasingly rely on such development processes [14]. Furthermore, with the emergence of architecture-based paradigms, we can begin to devise suitable architectures for intelligent systems [15]. The architecture based approach and software development processes go hand in hand offering many invaluable advantages such as incremental analysis, design, and testing. We believe, with the adoption of a synergistic development process (accounting for software, hardware, and bioware) combined with an appropriate architectural paradigm, we can incorporate, among other things, intelligence capabilities, metrics, and measurement methods in appropriate places.

⁴ We are currently writing these results for publication.
⁵ This is certainly a characteristic of e-commerce at internet speed.

3 ACKNOWLEDGEMENT

This research has been supported in part by NSF Next Generation Software (NGS) grant #EIA-9975050.

4 REFERENCES

- [1] Gigerenzer, G., P.M. Todd, (1999). Simple Heuristics That Make Us Smart, Oxford University Press.
- [2] Meystel, A.M., (2000). Simulation for Meaning Generation: Multiscale Coalitions of Autonomous Agents, in Discrete Event Modeling and Simulation Technologies: A Tapestry of Systems and AI-based Theories and Methodologies, Editors: H.S. Sarjoughian, F.E. Cellier, Springer Verlag.
- [3] Gautrais, J. T. Simon, (1998). Rate coding versus temporal order coding: a theoretical approach, Biosystems (48)1-3, pp. 57-65
- [4] Ruffin, V.R., J. Gautrais, A. Delorme, T. Simon, (1998). Face processing using one spike per neuron, Biosystems (48)1-3 pp. 229-239
- [5] Hans Moravec (1999). Rise of the Robots, Scientific American, August 1999, pp. 124-132
- [6] Zeigler, B.P., (1990). Object-Oriented Simulation with Hierarchical, Modular Models: Intelligent Agents and Endomorphic Systems., San Diego, CA: Academic Press
- [7] Fujimoto, R. (1998). "Time Management in the High-Level Architecture." Simulation 71(6): 388-400.
- [8] Zeigler, B.P., H. Praehofer, and T.G. Kim, (2000). Theory of Modeling and Simulation. 2ed, New York, NY: Academic Press
- [9] Zeigler, B.P., et al. (1998). The DEVS/HLA Distributed Simulation Environment and its Support for Predictive Filtering, ECE, The University of Arizona.
- [10] Zeigler, B.P., et. al. (1997). "The DEVS Environment for High-performance Modeling and Simulation." IEEE CS&E 4(3)
- [11] Zeigler, B.P. and H.S. Sarjoughian (1999). Support for Hierarchical Modular Component-based Model Construction in DEVS/HLA. Simulation Interoperability Workshop, Orlando, FL
- [12] Pressman, R. (1997). Software Engineering: A Practitioner's Approach, McGraw Hill
- [13] Booch G. (1996). Object Solutions: Managing the Object-Oriented Project. Menlo Park, CA, Addison-Wesley
- [14] Orfali, R., D. Harkey. (1997). The Essential Client/Server Survival Guide, John Wiley & Sons
- [15] Sarjoughian, H.S. and B.P. Zeigler (2001). "A Layered Modeling and Simulation Architecture for Agent-based System Development." IEEE Proceedings (to appear)

The Intelligence of an Entity

Copyright 2000©
Robby Glen Garner
Steven Boyd Henderson

Preface

Mimetic Synthesis is a new terminology that more accurately describes a programming methodology used to mimic human behavior in a computer such as a PC. Previous work in this field has been incorrectly categorized under various aspects of Artificial Intelligence (AI).

On Intelligence

Testing and quantifying intelligence is difficult at best, even if it's human intelligence. To Quote Tariq Samad from "Notes on Measuring Intelligence in Constructed Systems", "The difficulty of compressing the multifaceted nature of intelligence into one scalar quotient has led to proposals to consider intelligence not as one unitary quantity but as a collection of properties that are mutually incommensurable." Furthermore, one of the many lessons from a century of work on human intelligence is that we still don't really know what intelligence is.

Mimetic Entities

The early mimetic systems developed by Robby Garner are hierarchical in structure. This allows the "Mimetic Entity" to synthesize the combined behavior of subsystems into a unified presentation. This structure certainly suggests that one way to measure the intelligence of such machines is to review the hierarchical concepts it uses and the processes that contribute to the goals of the whole system.

One of the first hierarchical mimetic synthesizers was called Albert. This program combined the behavior of several methods that shared the same goal of simulating human conversation. Each method represents a separate strategy used to form the response to a human stimulus phrase.

The first method is based on a simple model of behavior, where conversation is represented by strings of (stimulus → response) nodes. The goal of this particular method is to find a match for the user's input stimulus in a database, and form the reply with the corresponding "response" from the database. If the first method is not successful, the program follows down the hierarchy from most specific method, to least specific.

The second method looks in a table of Boolean rules and attempts to fit a rule to the user's input. If a rule is satisfied, its corresponding response is used. The goal of this method is to satisfy a Boolean expression based on the user's input phrase.

And so on, the third method attempts to find a generalization about the user's input phrase using a "framed" template to determine a match. The goal of this method is to find a generalization that applies to the user's input phrase.

Then finally, if none of the other methods has succeeded, a final method selects a "new topic" from a pool of unused topics. The goal of this method is merely to make a response. (To change the subject)

So, one can see that the overall goal of simulating conversation is attempted by using a variety of strategies, all contributing to the main goal. The hierarchical structure ensures that the best possible response may be used.

It must be obvious that the performance of the mimetic entity with regards to simulating a conversation depends entirely on the performance of all of these various methods or subsystems. Yet it depends first and foremost on the person talking to it.

The Loebner Show

But what can we say about Albert's intelligence? None of the methods used are intelligent, so their "unified" representation is not intelligent. Albert may be perceived as intelligent by a human being as is evidenced by the 1998 Loebner Prize Contest, but the program is not in fact intelligent. <http://www.cs.flinders.edu.au/research/AI/LoebnerPrize/>

Then if we can know what intelligence is not, does that tell us what intelligence is?

No, because none of the competitors in the Loebner contest have exhibited intelligence. At best they exhibit a behavior which seems familiar to the user (judge), and some of them have used very clever means to achieve this. But the ingenuity of the programmer does not make the program intelligent.

One also has to agree that an imitation is not the same as the thing it imitates. Furthermore, some may object to things that are artificial for no other reason except that they are artificial. Yet if a thing works, does it matter why it works or what it is made from? Some people would say that if a thing is not really "intelligent" then it is an impostor, and therefore "dangerous." But if a tool performs a job according to specification, why is that less intelligent than if a human being had performed the same job?

By doing a job, there is at least one goal implied, and that is the completion of the job. If a computer completes the same job as a human in a smaller amount of time, we would say the computer has better performance, not better intelligence?

Human Intelligence

In dealing with other people, we assess their intelligence on a casual basis by observing their behavior, the things they say, their solutions to problems, or other factors, many of which are purely subjective.

Measuring machine intelligence would be much easier if people could agree on how to measure human intelligence!

So I think there is always a disparity between "perceived intelligence" and "actual intelligence", especially in evaluation of human intelligence. Intelligence is not solely performance, but is it possible to measure intelligence without also measuring a performance?

Sometimes a performance involves a great deal of preparation and training. If a man repeats the same sequence of behavior, practices it over and over until it can be done repetitively without thinking, is that intelligence?

Summary

The key to true intelligence is the ability of an entity to enlist strategy to accomplish its mission, not preconceived knowledge, or rote behavior.

Military confrontation is a good example according to R. Neil Bishop. "Time and time again, superior firepower and resources have been overcome by an inferior force with an intuitive strategy, which gave them a monumental advantage."

Also strategy is the key element needed to develop successful research techniques which, in pure science, may not even exist before the scientist begins. The strategy of obtaining and integrating knowledge is the key to reaching beyond what is presently known or understood.

The use of strategy applies not only to the highest level of abstraction, but is also evident in the "rank and file" subsystems that perform even the most basic tasks required by an entity as a whole. The strategy or algorithm employed by a programmer may be akin to "instinct" in some systems. Is instinctive behavior intelligent?