Proceedings of the Performance Metrics for Intelligent Systems (PerMIS) Workshop,
Gaithersburg, MD, August 13-15, 2002.

Metrics and Performance Measures for Intelligent Unmanned Ground Vehicles

James S. Albus
Senior NIST Fellow
Intelligent Systems Division
Manufacturing Engineering Laboratory
National Institute of Standards and Technology
Technology Administration
United States Department of Commerce

Abstract

Metrics and measures for physical phenomena are precisely defined and widely accepted. However, metrics and measures for intelligent systems are as yet vaguely defined and controversial. Even the definition of intelligence is not widely agreed upon.

A number of metrics and measures have been developed for measuring human performance in scholastic aptitude, athletic ability, and task performance. Some of these suggest metrics and performance measures for intelligent machine systems. An example of a set of performance measures for unmanned military scout vehicles is presented.

Keywords: metrics, performance measures, unmanned ground vehicles, intelligent systems

1. INTRODUCTION

Webster defines a metric as a standard of measurement. Examples include the meter (a standard of length), the kilogram (a standard of mass), the volt (electromotive force), the ampere (electric current), the second (time), and degree Celsius, Fahrenheit, or Kelvin (temperature.) A metric is what is used to make a measurement. For physical metrics there is wide agreement on how the metric is defined and how it can be applied to precisely measure a physical entity or temporal event.

However, when we come to a metric for intelligence, there is much less agreement and much less precision. There is not even agreement on what intelligence is, much less on how to measure it, or even what is the metric for measuring it. Almost any meeting can grind to a halt over the attempt to define intelligence. I don’t want that to happen here, so I am going to simply state my definition of intelligence, and move on.

Df: An intelligent system is a system with the ability to act appropriately in an uncertain environment

> where
> appropriate action is that which maximizes the likelihood of success in achieving or maintaining the highest level system goal. [Albus91]

The first thing to note about this definition is that intelligence has to do with a goal. Somewhere a goal is defined, and somehow the system accepts this goal as its own. The system then generates action that maximizes the likelihood that the goal with be achieved.

Note that the appropriate action may, or may not, be to head directly toward the goal. At the very lowest level, the proper action may depend not only on the current position relative to the goal, but on the current velocity and inertia of the system being controlled. At a higher level, a path planner may observe that the current vehicle position is in a cul-de-sac that blocks movement in the direction of the goal and plan a path away from the goal to escape the cul-de-sac.

Note also the reference to the "highest level" system goal. This implies a hierarchy of goals and subgoals with different planning horizons in time and space. At lower levels, goals are short-term, near-by, and high-resolution in time and space. At higher levels, goals are more distant and less precise in time and space.

Higher level goals may require that the system estimate the state of the world, gather information, build maps, plan routes, predict the future, imagine possible situations, weigh costs and benefits of alternative courses of action, and decide what behavior is most likely to achieve or maintain goals. For very high-level long-range goals such as rearing children, growing crops, or engaging in war, an intelligent system may need to make short term sacrifices, invent tools, develop weapons, and engage in deception.

2. LACK OF METRICS AND MEASURES

A major barrier to the development of intelligent systems is the lack of metrics and quantifiable measures of performance. There cannot be a science of intelligent systems without standard units of measure. To do science, you must be able to measure what you are doing and measure the results against some metric. This is something the field of AI, robotics, and intelligent systems has largely ignored. Most research results are in the form of demonstrations rather than experiments with data that that is quantitative and referenced against ground truth. There are few benchmarks or standardized tests wherein performance can be compared. That, of course is the subject of this workshop.

Perhaps the most common metric for human intelligence is the intelligence quotient (I.Q.) The average human I.Q. is arbitrarily defined as 100. But there is great controversy what I.Q. is and how it should be measured. There are, of course, many mental skills and abilities that can be measured. These include the ability to read and write, the ability to calculate with numbers, to reason with logic, to remember what was seen and heard, to perceive patterns, to understand relationships, and perform geometrical transformations. There are artistic skills and abilities. These include the ability to draw, paint, and sculpt, to sing and dance, to perform music, to compose poetry, to create or act out stories. There are manual skills and abilities to build or fix things, and athletic skills and abilities to compete in sports or fight in battles. In each of these areas, performance can be tested and scored.

Metrics for performance include speed (how fast?), precision (how accurate?), style (how graceful or well formed?), success/fail (criterion met?), effectiveness (desired result achieved?), efficiency (resources expended?), or cost/benefit (benefits worth the cost?)

Another type of performance metric is a benchmark. The performance of a system can be measured by comparing it against some benchmark performance. A benchmark may be an average over a population, or it may be a record of some kind, e.g., a world record, or a personal best.

One possible standard of measure is human performance. Human performance has been well defined and carefully calibrated in many areas. There are many existing measures of human performance, and human subjects are widely available. But how should this metric be applied? What should be measured? Perhaps the most fundamental measure is effectiveness in achieving goals. Certainly it is possible to measure timeliness. Cost, benefit, and risk are easily quantified for many tasks.

Typically a performance measure yields a score. The score may be some absolute quantity such as total points[1], or kilometers per hour, number of interventions per kilometer. Or the score may be a relative quantity such as order of finish in a race, or percentile in a distribution.

Sometimes the score takes into consideration the degree of difficulty of the performance. For example in competitive diving or figure skating, the performance score is multiplied by the degree of difficulty to decide the winner of a competition. In other cases, effort is made to assure that all competitors experience the same degree of difficulty – a so-called "level playing field." For example, in basketball and football games the teams switch goals at half-time. In races, all competitors are required to cover the same distance. In competitions where a level playing field cannot be achieved, there may be a preliminary competition to decide who gets the advantage. For example in automobile racing, qualifying time determines the starting line up. In tournaments, preliminary competition determines who meets the weakest competitor.

For autonomous driving, the degree of difficulty depends on the environment. On-road driving is more difficult on crowded streets and at intersections than on deserted roads and empty streets. The level of difficulty of off-road driving depends on the terrain and ground cover. It also depends on the density of obstacles such as ditches, trees, and rocks, and whether obstacles are hidden beneath tall grass or dense weeds.

In many cases, it is necessary to take into account the amount of training and preparation that have preceded the testing process. To accommodate these variations, different classes of competition may be established. Thus there are many issues with regard to how performance measurements should be made and how the results should be scored.

[1] where total points = points-per-goal x number of goals

Finally, there is the issue of what the score means. Typically the competitor with the most total points wins,[2] and rank order is determined by the number of points scored. However, not all competitions mean the same, and all wins or losses are not the same. Winning a pre-season game is not the same as winning the Super Bowl. In measuring the performance of intelligent systems, not all tests are equal. Passing a routine drivers test is not the same as qualifying for the "Indianapolis 500."

3. PERFORMANCE MEASURES

To address the issues of performance measures for intelligent systems, NIST has begun work in three areas:

1) a test course for search and rescue robots,
2) a measurement procedure for evaluating run-off-road detectors, and
3) a set of performance measures for autonomous driving.

The NIST test course for urban search and rescue addresses the problems of searching for human victims in buildings that have collapsed because of earth quakes, terrorist attacks, or other disasters. The USAR test course has been used in several AAAI and RoboCup competitions around the world. [Jacoff et al.00, 01, 02] This work is sponsored by the DARPA Mobile Autonomous Robot Software (MARS) program.

The NIST measurement procedure for run-off-road detectors addresses the problem of evaluating commercial products for effectiveness in determining when a vehicle is in danger of running off the road and warning the driver in time to prevent an accident. [Szabo et al.99] This work is sponsored by the Department of Transportation Highway Safety administration.

NIST work on metrics, performance measures, and standard reference data for autonomous driving addresses both off-road and on-road applications. This work is sponsored by the Army Research Lab Demo III Experimental Unmanned Vehicle (XUV) program. The Army is interested in measuring the state of readiness of autonomous driving technology for unmanned military vehicles. [Bornstein02] Specifically, the tests are designed to determine whether the Demo III XUVs have achieved technology readiness level six (TRL-6). TRL-6 requires that a prototype be demonstrated in a relevant environment.

For autonomous mobility, we assume that the relevant environment includes driving off-road through tall grass, weeds, and brush; through woods and fields, in desert and mountain terrain. The relevant environment also includes driving on-roads of all types including overgrown dirt trails, gravel roads, paved rural roads and highways, as well as and urban paved streets and alleys that may contain piles of rubble, burning tires, and abandoned vehicles. Relevant environmental conditions include day, night, rain, dry, dust, smoke, mud, and possibly snow and ice. It will not include the ability to

[2] except in golf where the competitor with the lowest score wins

autonomously cope with on-coming traffic, pedestrians, animals, moving vehicles, intersections, traffic signals, or road signs.

3.1 TRL-6 Test Procedures

The TRL-6 tests will proceed as follows:
For a chosen set of missions in a variety of environments (woods, fields, roads, trails, urban, desert, mountains) and a variety of conditions (day, night, dry, wet, snow, mud, dust, smoke):

- A manned scout vehicle will perform an assigned mission and its performance will be measured and scored for military effectiveness in terms of mission success, timeliness, resource expenditure, risk, and probable human casualties.
- A robot scout vehicle will perform the same mission and its performance will be measured and scored by the same criteria.
- The performance and score of the robot scout vehicle will be compared against that of the manned scout vehicle.
- A terrain characterization vehicle will exactly retrace the routes traversed by both manned and robot vehicles and will characterize the difficulty of the terrain covered in terms of slope, roughness, soil mechanics, and ground cover.
- Terrain characterization will help to define the operational envelope within which the robot vehicles can be used effectively.

3.2 Terrain Characterization

At least one way to measure difficulty is to measure the surface attributes of the terrain. The following is a set of terrain measurement techniques proposed for the coming TRL-6 experiments.

There will be a baseline and advanced set of terrain characterization measurements.

1. The baseline measurements will consist of one or more human observers riding in a HMMWV and subjectively scoring the difficulty of the terrain.
2. A more advanced scenario will include roughness measurements from an inertial navigation system, a TV camera, and an instrumented bumper on the XUV. These measurements will be compared with similar measurements made by similar sensors on the manned scout vehicle.
3. Still more advanced scenarios will include measurements made from the terrain characterization vehicle by high precision high-resolution LADAR .
4. The most advanced scenarios will include measurements of soil mechanics made from the terrain characterization vehicle.
5. If necessary, additional terrain characterization data will be obtained from overflights using airborne stereo cameras or LADAR scanners.

The terrain will be characterized by the following method:

1. A terrain characterization vehicle (HMMWV) will be driven over the exact paths traveled by the manned scout vehicle and the XUV during their respective missions.

2. The paths chosen by both vehicles will be scanned at regular intervals by high resolution LADAR cameras. These range images will then be registered with color images from color cameras. Data from an INS system will measure accelerations produced by bouncing over the terrain. Data from an instrumented bumper will measure the strength of the vegetation being driven through. Instrumentation to measure soil mechanics may also be carried by the terrain characterization vehicle.
3. The point clouds from the high resolution LADARs will be stitched together and the terrain will be characterized in terms of slope, roughness, obstacles, and ground cover. The conditions during the respective missions will be characterized by one or more of terrain characterization measurements described above.
4. Missions will be run under a variety of conditions including different times of day and night, different lighting, wet and dry, and different amounts of smoke and dust. The terrain attributes and conditions will be combined to provide a measure of difficulty for each path.
5. The two paths selected by the manned and robot vehicles will be scored and compared to determine which was the “better” path for the mission.

The difficulty of the terrain will also be analyzed to characterize the operational envelop within which the robot vehicle can be expected to reliably perform.

3. 3 Scout Vehicle Test Scenarios

The performance of the manned scout vehicle will be measured by the following procedure:

1. The Test Director will give a human commander a typical scout mission to be accomplished. The human commander will issue orders to the driver of a manned scout vehicle (a HMMWV) via a radio operator. The commander will be located in a HMMWV following the scout vehicle at a prescribed distance. The commander will have a map display overlaid with the manned scout location, the command vehicle location, and mission objectives. The commander may or may not have visual contact with the scout vehicle depending on the terrain and the separation distance between commander and manned scout vehicle.
2. A human driver in the manned scout vehicle will drive the scout vehicle in a tactical manner to accomplish the assigned mission objectives.
3. The manned scout vehicle will be scored by conventional methods used for evaluating human scout performance.
4. The number and type of conversations between the scout vehicle, the radio operator, and the commander will be measured.
5. The work load on the radio operator and commander will be measured.
6. The bandwidth and total amount of communications between manned vehicle, the radio operator, and the commander will be measured.
7. The behavior of the human driver, commander, and ratio operator will be video taped
8. A panoramic camera mounted on the manned scout vehicle will be used to record the scenes encountered by the human driver.

The performance of the robot scout vehicle (a XUV) will be measured by the following procedure:

1. The Test Director will give the human commander of the XUV the same mission as given to the commander of the manned scout vehicle. The human commander will issue orders to the XUV via a radio operator. The commander will be located in a HMMWV following the XUV at a prescribed distance. The commander will have a map display with the XUV location, the command vehicle location, and mission objectives. The commander may or may not have visual contact with the XUV depending on the terrain and the separation distance between commander and XUV.
2. The Demo III autonomous mobility system will drive the XUV to accomplish its mission objectives. The XUV may, or may not, choose the same path over the terrain as the manned scout vehicle. Either the commander or radio operator can provide intermediate way points to help robot get to goal point. Only the radio operator can teleoperate the robot.
3. The XUV performance will be scored by the same methods used for evaluating the manned scout vehicle.
4. The number of interventions by the radio operator will be measured and the type of interventions will be classified and analyzed.
5. The work load on the radio operator and commander will be measured
6. The bandwidth and volume of communications between XUV, the radio operator, and the commander will be measured.
7. A trace will be kept of critical state variables and world model representations during and prior to operator interventions.
8. The behavior of the XUV, the radio operator, and the commander will be video taped
9. A panoramic camera mounted on the XUV will be used to record the scenes encountered by the XUV.

4. SUMMARY AND CONCLUSIONS

Metrics and measures for physical phenomena are precisely defined and widely accepted. However, metrics and measures for intelligent systems are as yet vaguely defined and controversial. Even the definition of intelligence is not widely agreed upon.

There are a number of metrics and measures that have been developed for measuring human performance in scholastic aptitude, athletic ability, and task performance. It is suggested that these may provide guidelines for developing metrics and performance measures for intelligent machine systems. An example of a set of performance measures for unmanned military scout vehicles is presented.

5. REFERENCES

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

Bornstein, J. A. (2002) “U.S. Army ground robotics research program”, *Proceedings of SPIE Aerosense Conference Vol. 4715,* Orlando, Florida, 1-5 April, 2002

Jacoff, A., Messina, E., Evans, J. (2000) A Standard Test Course for Urban Search and Rescue Robots, *Proceedings of the Performance Metrics For Intelligent Systems Workshop*, National Institute of Standards and Technology, Gaithersburg, MD, August 14-16

Jacoff, A., Messina, E., Evans, J. (2001) Experiences in Deploying Test Arenas for Autonomous Mobile Robots, *Proceedings of the Performance Metrics For Intelligent Systems (PerMIS) Workshop*, in association with IEEE CCA and ISIC, Mexico City, Mexico, Sept 4, 2001

Jacoff, A., Messina, E., Evans, J. (2002) Performance Evaluation of Autonomous Mobile Robots, *Industrial Robot 29:3*

Szabo, S., Murphy, K., Juberts, M. (1999) The AUTONAV/DOT Project: Baseline Measurement System for Evaluation of Roadway Departure Warning System, *NISTIR 6300,* National Institute of Standards and Technology, Gaithersburg, MD, March