Hypothesis Testing for Complex Agents

Joanna Bryson, Will Lowe† and Lynn Andrea Stein
MIT AI Lab
Cambridge, MA 02139
joanna@ai.mit.edu, las@ai.mit.edu

†Tufts University Center for Cognitive Studies
Medford, MA 02155
wlowe02@tufts.edu

Abstract

As agents approach animal-like complexity, evaluating them becomes as difficult as evaluating animals. This paper describes the application of techniques for characterizing animal behavior to the evaluation of complex agents. We describe the conditions that lead to the behavioral variability that requires experimental methods. We then review the state of the art in psychological experimental design and analysis, and show its application to complex agents. We also discuss a specific methodological concern of agent research: how the robots versus simulations debate interacts with statistical evaluation. Finally, we make a specific proposal for facilitating the use of scientific method. We propose the creation of a web site that functions as a repository for platforms suitable for statistical testing, for results determined on those platforms, and for the agents that have generated those results.

Keywords: *agent performance, complex systems, behavioral indeterminacy, replicability, experimental design, subjective metrics, benchmarks, simulations, reliability.*

1. Introduction

Humanoid intelligence is a complex skill, with many interacting components and concerns. Unless they are in an exceptional, highly constrained situation, intelligent agents can never be certain they are expressing the best possible behavior for the current circumstance. This is because the problem of choosing an ordering of actions is combinatorially explosive [9]. Consequently, for scientists or engineers evaluating the behavior of an agent, it is generally impossible to ascertain whether a behavior is optimal for that agent. Albus [2] defines intelligence as "the ability of a system to act appropriately in an uncertain environment, where appropriate action is that which increases the probability of success." Systems of such complexity are rarely amenable to proof-theoretic techniques [26]. In general, the only means to judge an increase in probability is to run statistical tests over an appropriately sized sample of the agent's behavior.

Computational systems, in contrast, are traditionally evaluated based on their *final* results and/or on their resource utilization [29]. The historical definition of computational process (c.f. Babbage, Turing, von Neumann) is modeled on mathematical calculation, and its validity is measured in terms of its ultimate product. If the output is correct — if the correct value is calculated — then the computation is deemed correct as well. More recent descriptions [e.g. 11] have added an assessment of the time, space, processor, and other resource utilization, so that a computation is only deemed correct if it calculates the appropriate value within some resource constraints.

This characterization of computation is less applicable when it comes to particular operating systems and other real-time computational systems. These systems have no final result, no end point summarizing their work. Instead, they must be evaluated in terms of ongoing behavior. Guarantees, where they exist, take the form of performance constraints and temporal invariants. Although formal analysis of correctness plays a role even in these systems, performance testing, including benchmarking, is an essential part of the evaluation criteria for this kind of computational system.

Computational agent design owes much to computer science. But the computationalist's tendency to evaluate in terms of ultimate product is as inappropriate for computational agents as it is for operating systems. Instead, metrics must be devised in terms of ongoing behavior, performance rather than finitary result. But what is the analog to benchmarking when the tasks are under-specified, ill-defined, and subject to interpretation and observer judgment?

In this paper, we will examine issues of running such evaluations for complex agents. By *complex agents* we mean autonomous agents such as robots or VR characters capable of emulating humanoid or at least vertebrate intelligence. We will discuss hypothesis testing, including the statistical controversies that have lead to the recent revisions in the standard experimental analysis endorsed by the American Psychological Association. We will also discuss recent advances in methodologies for establishing quantitative metrics for matters of human judgment, such as whether one sentence is more or less grammatical than another, or an anecdote is more or less appropriate. We propose a means to facilitate hypothesis testing between groups: a simulation server running a number of benchmark tests.

2. Motivation: Sources of Uncertainty

Although there is certainly a role for using formal methods in comparing agent architectures [e.g. 8, 6], what we as agent designers are ultimately interested in is comparing the resulting *behavior* of our agents. Given the numerous complex sources of indeterminacy in this behavior, such comparison requires the application of the same kind of experimental methodology that has been developed by psychology to address similar problems. In this section we review some of the sources of this indeterminacy; in the next we will review analytic approaches for addressing them.

The first source of indeterminacy is described above: The combinatorial complexity of most decision problems makes absolute optimality an impractical target. Thus even if there is a single unique optimal sequence of actions, in most situations we cannot expect an agent to find it. Consequently, we will expect a range of agents to have a range of suboptimal behaviors, and must find a way of comparing these.

The next source of indeterminacy is the environment. Many agents must attempt to maintain or achieve multiple, possibly even contradictory goals. These goals are often themselves uncertain. For example, the difficulty of eating is dependent on the supply of food, which may in turn be dependent on situations unknowable to the agent, whether these be weather patterns, the presence or absence of other competing agents, or in human societies, local holidays disrupting normal shopping. Thus in evaluating the general efficacy of an agent's behavior, we would need a large number of samples across a range of environmental circumstances.

Another possible source of indeterminacy is the development of agents. As engineers, we are not really interested in evaluating a single agent, but rather in improving the state-of-the-art in agent design. In this case, we are really interested in what approaches are most likely to produce successful agents. This involves uncertainty across development efforts, complicated by individual differences between developers. Many results contending the superiority or optimality of a particular theory of intelligence may simply reflect effective design by the practitioners of that theory [e.g. 7].

Finally, the emphasis of this workshop is on natural, human-like behavior. Humans are highly social animals, and social acceptability is an important criteria for intelligent agents. However, sociability is not a binary attribute: it varies in degrees. Further, a single form of behavior may be considered more or less social by the criteria of various societies. Evaluations of systems by such criteria requires measurement over a population of judges.

3. Current Approaches to Hypothesis Testing

The previous section presented a number of challenges to the evaluation of complex, humanoid agent building techniques. In this section we review methodologies used by psychology — the evaluation of human agents — that are available to address these challenges.

Although it is obvious that comparing two systems requires testing, the less obvious issues are how many tests need to be run and what statistical analysis needs to be used in order to answer these questions. In this section we describe three increasingly common problems in Artificial Intelligence and discuss a set of experimental techniques from the behavioral sciences that can be used to address them.

The first problem is variability in results: We need to know whether performance differences that arise over test replications can be ascribed to varying levels of a system's ability or to variation in lighting conditions, choice of training data, starting position, or some other or some other external (and therefore uninteresting) source. Psychology uses statistical techniques such as the Analysis of Variance (ANOVA) to address these issues. The second problem is of disentangling complex and unexpected interactions between subparts of a complex system. This can also be addressed using ANOVA coupled with factorial experimental design. The third problem is that of rigorously and meaningfully evaluating inherently subjective data. Since many psychology experiments investigate inherently subjective matters, the field has developed a set of techniques that will be of use to artificial agent designers as well. The next three sections describe these solutions in more detail.

3.1 Variability in Results

The problem of comparing performance variability due to differences in ability and variability due to extraneous factors is ubiquitous in psychology. It is dealt with by procedures known collectively as Analysis of Variance or ANOVA.

3.1.1 Standard ANOVA

In a typical experimental design for comparing performance, K systems are tested N times each. If the variation in performance between the K systems outweighs the variability among each system's N runs, then the system performances are said to be *significantly different*. We then examine the systems pairwise to get information about ordering. The ANOVA allows us to infer that e.g. although there are differences overall between the K=4 systems (i.e. some are better than others), the performance difference between 3 and 4 is reliable, whereas the difference between 1 and 2 is not reliable because it is outweighed by the amount of extraneous variation across the N tests. In this case, although 1 may perform on average better than 2, this does not imply that it is actually better on the task. If the experiment were repeated then 2 would have reasonable chance of performing on average the same as 1, or even better.

The notion of *reasonable chance* used above is the essence of the concept of significant difference. System 3 is on average better than 4 in this experiment and the ANOVA tells us that performances are significantly different at the .05 level (expressed as $p<.05$). This means that in an infinite series of experimental replications, if 3 is in fact exactly *as good* as 4, i.e. there is no genuine performance difference, then the probability of getting a performance difference as large or larger than the one observed in this experiment is 0.05. The smaller this probability becomes, the more reliable the difference is. In contrast, the fact that the average performances of 1 and 2 are not significantly different means their ordering in this experiment is not reliable because there is a more than 0.05 probability that the ordering would not be preserved in a replication.

Notice that hypothesis testing using ANOVA does not *guarantee* an ordering, it presents probabilities that each part of the ordering is reliable. This is a fundamental difference between experimental evidence and proof. Scientific method increases the probability that hypotheses are correct but it does not demonstrate them with complete certainty.

The binary output of hypothesis tests (significant difference versus no significant difference) and its probability is an unnecessarily large loss of information. The American Psychological Association have consequently recently moved to emphasize confidence intervals over simple hypothesis testing. A *confidence interval* is a range, centered on the observed difference, that in the hypothetical replications will contain the true performance value some large percentage, say 95%, of the time. In the example above, each system has a 95% confidence interval, or *error bar*, centered on its average performance with width determined by the amount of variability between runs. When two intervals overlap, there is a significant probability that a replication will not preserve the current ordering among the averages and we can conclude that the corresponding performance difference is unreliable. This method gives the same result as simple hypothesis testing above — the performances are not significantly different — but is much more informative: confidence intervals give an idea about how much variability there is in the data itself and yield a useful graphical representation of analytical results.

3.1.2 Alternative Approaches to Analysis

Stating confidence intervals is more informative than simple significance judgments. However, it also relies on an hypothetical infinity of replications of an experiment. This aspect of classical statistical inference is a result of assuming that the true difference in performance is fixed and the observed data is a random quantity. Alternatively, in Bayesian analysis the difference is considered uncertain and is modeled as a random variable whereas the results are fixed because they have already been observed [5]. The result is a probability distribution over values of the true difference. To summarize the distribution an interval containing 95% of the probability mass can be quoted. This takes the same form as a confidence interval, except that its interpretation is much simpler: Given the observed results, the probability that the true difference is in the interval is 0.95, so if the interval contains 0, there is a high probability that there is no real performance difference between systems.

The Bayesian approach makes no use of hypothetical experimental replications and is more naturally extended to deal with complicated experimental designs. On the other hand, it does require an initial estimate (or prior distribution) for the probabilities of various values of the performance difference before seeing test data. There is much controversy about which of these approaches is more appropriate. In the context of AI however, we need not take a stand on this issue. The two approaches answer different questions, and for our purposes the questions answered by classical statistics are of considerable interest. Unlike many of the natural sciences, the performance of AI systems over multiple replications is not only accessible, but of particular interest. To the extent we are engineers, AI researchers must be interested in reliability and replicability of results.

3.2 Testing for Interacting Components

Many unpleasant software surprises arise from unexpected interactions between components. Unfortunately, in a complex system it is typically infeasible to discover the nature of interactions analytically in advance. Consequently *factorial experimental design* is an important empirical tool.

As an example, assume that we can make two changes A and B to a system. We could compare the performance of the system with A to the same system without it, using the ANOVA methods above, and then do the same for B. But when building a complex system it is essential to also know how A and B affect performance together. Separate testing will never reveal, for example, that adding A generates a performance improvement only when B is present and not otherwise. This is referred to as an *interaction* between A and B, and can be dealt with by testing all combinations of system additions, leading to a factorial experiment. Factorial experiments are analyzed using simple extensions to ANOVA that test for significant interactions as well as simple performance differences. Factorial ANOVA methods are described in any introductory statistics textbook [e.g 23].

In the discussion above we have implicitly assumed that differences in performance can be modeled as continuous quantities, such as distance traveled, length of conversation or number of correct answers. When the final performance measure is discrete, e.g. success or failure, then *logistic regression* [1, ch.4] is a useful way to examine the effects of additions or manipulations on the system's success rate. Information about the effects of arbitrary numbers of additions, both individually and in interaction, is available using this method, just as in the factorial ANOVA. Logistic regression also gives a quantitative estimate of *how much* the probability of success changes with various additions to the system, which gives an idea of the importance of each change.

3.3 Quantifying Inherently Subjective Data

Often performance evaluation involves judgments or ratings from human subjects. Clearly it is not enough that one subject judges an AI conversation to be lifelike because we do not know how typical that subject is, and how robust their opinion is. It would be better to choose a larger sample of raters, and to check that their judgments are reliable. When ratings are discrete (good, bad) or ordinal (terrible, bad, ok, good, excellent) then Kappa [22] is a measure of between-rater agreement that varies from 1 (perfect agreement) to -1 (chance levels of agreement). For judgments of continuous quantities the intraclass correlation coefficient [13] performs the same task.

However, such discrete classifications are often clumsy. Because a rating system is itself subjective, the extra variance added by difference in interpretation of a category can lose correlations between subjects that actually agree on the relative validity or likeability of two systems. Further, we would really prefer in many circumstances to have a continuous range of difference values. Such results can be provided by *magnitude estimation*, a technique from psychophysics. For example, Bard *et al.* [4] have recently introduced the use of magnitude estimation to allow subjects to judge the acceptability of sentences which have varying degrees of syntactic propriety. In a magnitude estimation task, each subject is asked to assign an arbitrary number as a value for the first example they see. For each subsequent example, the subject need only say how much more or less acceptable it is, with reference to the previous value, e.g. twice as acceptable, half as acceptable and so on. This allows subjects to pick a scale they feel comfortable with manipulating, yet gives the experimenter a generally useful metric. For example, in Bard *et al.*'s work, a subject might give the first sentence an 8, the next a 4, the following a 32 — the experimenter records 1s, .5s and 4s respectively. This method has been shown to reduce the number of judgments necessary to get very reliable and accurate estimates of acceptability, relative to other methods.

Bard *et al.* manipulate the sentences themselves, but it is clear that magnitude estimation can equally well be used to get fine-grained judgments about how natural the output of a natural language processing (NLP) system is, and the degree to which this is improved by adding new components. Nor is the method limited to linguistic judgments, for it should be equally effective for evaluating ease of use for teaching software, the psychological realism of virtual agents or the comprehensibility of output for theorem proving machinery.

4. Environments for Hypothesis Testing: Robots and Simulations

As the previous sections indicate, one of the main attributes of statistically valid comparisons is a large number of experimental trials. Further, these experimental conditions should be easily replicable and extendible by other laboratories. In Section 5. we propose that a good way to facilitate such research is to create a web location dedicated to providing source code and statistics for comparative evaluations over a number of different benchmark tasks. This has approach has proven useful in neural network research, and should also be useful for complex agents. However, it flies in the face of one of the best-known hypotheses of complex agent research: that good experimental method requires the use of robots. Consequently, we will first provide an updated examination of this claim.

4.1 Arguments Against Simulation

Simulation is an attractive research environment because it is easy to maintain valid controls, and to execute large numbers of replications across a number of machines. However, there have been a number of important criticisms leveled against this approach.

A Simulations never replicate the full complexity of the real world. In choosing how to build a simulation, the researcher first determines the 'real' nature of the problem to be solved. Of course, the precise nature of a problem largely determines its solution. Consequently, simulations are not valid for truly complex agents, because they do not test the complete range of problems a natural or embodied agent would face.

B If a simulation truly were to be as complicated as the real world, then building it would cost more time and effort than can be managed. It is cheaper and more efficient to build a robot, and allow it to interact with the real world. This argument assumes one of basic hypotheses of the behavior-based approach to AI [3], that intelligence is by its nature simple and its apparent complexity only reflects the complexity of the world it reacts to. Consequently, spending resources constructing the more complicated side of the system is both irrational and unlikely to be successful.

C When researchers build their own simulations, they may deceive either themselves or others as to the validity or complexity of the agents that operate in it. Since both the problem and the solution are under control of the researcher, it is difficult to be certain that neither unconscious nor deliberate bias has entered into the experiments. In contrast, a robot is considered to be clear demonstrations of autonomous artifact; its achievements cannot be doubted, because it inhabits the same problem space we do.

4.2 *Are Robots Better than Simulations?*

These arguments have led to the wide-spread adoption of the autonomous robot as a research platform, despite the known problems with the platform [16]. These problems reduce essentially to the fact that robots are extremely costly. Although their popularity has funded enough research and mass production to reduce the initial cost of purchase or construction, they are still relatively expensive in terms of researcher or technician time for programming, maintenance, and experimental procedures. This has not prevented some researchers from conducting rigorous experimental work on robot platforms [see e.g. 10, 25]. However, the difficulty of such procedures adds urgency to the question of the validity of experiments in simulation.

This difficulty has been reduced somewhat by the advent of smaller, more robust, and cheaper mass-produced robot platforms. However, these platforms still fall prey to a second problem: mobile robots do not necessarily address the criticisms leveled above against simulations better than simulations do. There are two reasons for this: the need for simplicity and reliability in robots, and the growing sophistication of simulations.

The constraints of finance, technological expertise and researchers' time combine to make it extremely unlikely that a robot will operate either with perception anything near as rich as that of a real animal, nor with actuation having anything like the flexibility or precision of even the simplest animals. Meanwhile, the problem of designing simulations with predictive value for robot performance has been recognized and addressed as a research issue [e.g. 18]. All major research robot manufacturers now distribute simulators with their hardware. In the case of Khepera, the robot most used by researchers running experiments requiring large numbers of trials, the pressure to provide an acceptable simulator seems to have not only resulted in an improved simulator, but also a simplified robot, thus making results on the two platforms nearly identical. Clearly this similarity of results either validates the use of the Khepera simulator, or invalidates the use of the robot.

When a simulator is produced independent of any particular theory of AI as a general test platform, it defeats much of the objection raised in charges *A* and *C* above, that a simulator is biased towards a particular problem, or providing a particular set of results. In fact, complaint *C* is particularly invalid as a reason to prefer robotics. Experimental results provided on simulations can be replicated precisely in other laboratories. Consequently, they are generally *more easily* tested and confirmed than those collected on robots. To the extent that a simulation is created for and possibly by a community — as a single effort resulting in a platform for unlimited numbers of experiments by laboratories world-wide, that simulation also has some hope of overcoming argument *B*.

This gross increase in the complexity of simulations has particularly true of two platforms. First, the simulator developed for the simulation league in the RoboCup soccer competition has proven enormously successful. Although competition also takes place on robots, to date the simulator league provides far more "realistic" soccer games in terms of allowing the demonstration of teamwork between the players and flexible offensive and defensive strategies [21, 19]. This success has encouraged the RoboCup organization to tackle an even more complex simulator designed to replicate catastrophic disasters in urban settings [20]. This simulator is intended to be sufficiently realistic as to eventually allow for swapping in real-time sensory data from disaster situations, in order to allow disaster relief to monitor and coordinate both human and robotic rescue efforts.

The second platform is also independently motivated to provide the full complexity of the real world. This is the commercial arena of virtual reality (VR), which provides a simulated environment with very practical and demanding constraints which cannot easily be overlooked. Users of virtual reality bring expectations from ordinary life to the system, and any agent in the system is harshly criticized when it fails to provide adequately realistic behavior. Thórisson [30] demonstrates that users evaluate a humanoid avatar with which they have held a conversation as much more intelligent if it provides back-channel feedback, such as eyebrow flashes and hand gestures, than when it simply generates and interprets language. Similarly Sengers [27] reviews evidence that users cannot become engaged by VR creatures operating with overly reactive architectures, because the agents do not spend sufficient time telegraphing their intentions or deliberations. Such constraints have often been overlooked in robotics.

In contrast, robots which must be supported in a single lab with limited technical resources are likely to deal with far simpler tasks. Robots may face far fewer conflicting goals, lower time-related conflicts or expectations, and even fewer options for actuation. Although robots still tend to have more natural perceptual problems than simulated or VR agents, even these are now increasingly being addressed with reliable but unnatural sensors such as laser range finders.

4.3 *Roles for Robots and Simulations*

Robots are still a highly desirable research platform. They provide complete systems, requiring the integration of many forms of intelligence. Many of the problems they need to solve are closely related to animal's problems, such as perception and navigation. In virtual reality, perfect perception is normally provided, but motion often has added complication over that in the real world. Depending on the quality of the individual virtual reality platform, an agent may have to deliberately not pass through other objects or to intentionally behave as if it were affected by gravity or air resistance. Even in the constantly improving RoboCup soccer simulator, there are outstanding difficulties in simulating important parts of the game, such as the goalkeeper's ability to kick over opposing team members (currently compensated for by allowing the keeper to "warp" to any point in the goal box instantaneously when already holding the ball.)

Robots being embodied in the real world are still probably the best way to enforce certain forms of honesty on a researcher. A mistake cannot be recovered from if it damages the robot, an action once executed cannot be revoked. Though this is also true of some simulations [e.g. 31], particularly in the case of younger students, these constraints are better brought home on a robot, as it becomes more apparent why one can't 'cheat.' Finally, building intelligent robots is a valid end in itself. Commercial intelligent robots are beginning to prove very useful in care-taking and entertainment, and may soon prove useful in areas such as construction and agriculture. In the meantime robots are highly useful in the laboratory for stirring interest and enthusiasm in students, the press and funding agencies. However, given the arguments above, we conclude that the use of robots as experimental platforms is neither necessary nor sufficient in providing evidence about complex agent intelligence. Robots, like simulations, must be used in combination with rigorous experimental technique, and even so can only provide evidence, not conclusive proof, of agent hypotheses.

In summary, neither robots nor simulation can provide a single, ultimate research platform. But then, neither can any other single research platform or strategy [15]. While not denying that intelligence is often highly situated and specialized [14, 17], to make a general claim about agent methodology requires a wide diversity of tasks. Preference in platforms should be given to those on which multiple competing hypotheses can be tested and evaluated, whether by qualitative judgments such as the preference of a large number of users, or by discrete quantifiable goals to be met, such as a genetic fitness function, or the score of a soccer game.

5. Coordinating Hypothesis Testing

Whether there can be general solutions to problems of intelligence is an empirical matter that has already been tested in some domains. For neural networks and other machine learning methods, the UCI Machine Learning Repository holds a large collection of benchmark learning tasks. Besting these benchmarks is not a necessary requirement for the publication of a new algorithm, but showing a respectable performance on them improves the reception of new contributions. Essentially, benchmarks are one indication for both researchers and reviewers of when an innovation is likely to be of interest.

Further, Neal and colleagues at the University of Toronto have constructed DELVE [24], a unified software framework for benchmarking machine learning methods. DELVE contains a large number of benchmark data sets, details of various machine learning techniques, currently mostly neural networks and Gaussian Processes, and statistical summaries of their performance on each task. One of the most important requirements is that each method is described in enough detail that it could be implemented by another researcher and would obtain a similar performance on the tasks. This ensures that the mundane but essential decisions that are an essential part of many learning algorithms (e.g. setting weight decay parameters, choosing k in k-nearest-neighbor rules) are not lost.

We propose a complex agent comparison server or web site, to be at least partially modeled on DELVE. This site should allow for the rating of both agent approaches and comparison environments, thus encouraging and facilitating research in both fields. It could also be annotated for educational purposes, indicating challenges and environments well suited to school, undergraduate, and graduate course projects. Such a site might provide multiple indices, such as:

- Environments, ranked by number and/or diversity of participants.
- Agent architectures (e.g. Soar, Behavior-Based AI). This should also allow for the petition for new categories.
- Contestants and/or contesting labs or research groups . This allows researchers interested in a particular approach to see any related work. Ranked by the number and/or diversity of environments.

Here are some examples of already existent platforms which might be included on the server:

- RoboCup [21, 19].
- Khepera robot competitions. Both of these two suggestions provide simulations as well as organized robotic competitions. They test learning and perception as well as planning or action selection.
- Tile World and Truck World, designed as complex planning domains. [15]
- Tyrrell's Simulated Environment [31] designed to test action-selection and goal management.
- Chess.
- An analog Turing Test, using magnitude estimation to compare dialog systems.

In addition, there are at least two software environments designed specifically to allow testing and comparison of a number of different architectures, though they contain no specific experimental situations as currently developed. These environments are Cogent [12] and the Sim_agent Toolkit [28].

6. Conclusion

To summarize, we believe that as agents approach the goal of being psychologically realistic and relevant, their evaluation will require the techniques that have been developed in the psychological sciences. This evaluation is critical in providing a gradient as we search for the right sorts of techniques to build complex agents. The techniques of hypothesis testing have been refined to describe truly complex agents. However, these are scientific techniques, not proofs. They do not give us certain answers, only more information. We believe many of the criticisms of benchmark testing made in the past failed to properly acknowledge this feature of experimentation. We should trust increased probability, rather than proof-theoretic guarantees. The more people perform tests across competing hypotheses, the more likely we will be to achieve our research goals, whether they are engineering complex, social agents, or understanding the nature of intelligence.

Acknowledgments

The authors would like to acknowledge early discussions with Brendan McGonigle and Ulrich Nehmzow on this topic.

References

A. Agresti. *Categorical Data Analysis*. John Wiley and Sons, 1990.

J. S. Albus. Outline for a theory of intelligence. *IEEE Transactions on Systems, Man and Cybernetics*, 21(3):473–509, 1991.

Ronald C. Arkin. *Behavior-Based Robotics*. MIT Press, Cambridge, MA, 1998.

E. Bard, D. Robertson, and A. Sorace. Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68, 1996.

G. E. P. Box and G. C. Tiao. *Bayesian inference in statistical analysis*. Addison-Wesley, Reading, Massachusetts, 1993.

Joanna Bryson. Cross-paradigm analysis of autonomous agent architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, 12(2):165–190, 2000.

Joanna Bryson. Hierarchy and sequence vs. full parallelism in reactive action selection architectures. In *From Animals to Animats 6 (SAB00)*, Cambridge, MA, 2000. MIT Press.

Joanna Bryson and Lynn Andrea Stein. Architectures and idioms: Making progress in agent design. In *The Seventh International Workshop on Agent Theories, Architectures, and Languages (ATAL2000)*, 2000. to be presented July 2000.

David Chapman. Planning for conjunctive goals. *Artificial Intelligence*, 32:333–378, 1987.

David Cliff, Philip Husbands, and Inman Harvey. Explorations in evolutionary robotics. *Adaptive Behavior*, 2(1):71–108, 1993.

S. A. Cook. The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on the THeory of Computing*, pages 151–158, New York, 1971. Association for Computing Machinery.

R. Cooper, P. Yule, J. Fox, and D. Sutton. COGENT: An environment for the development of cognitive models. In U. Schmid, J. F. Krems, and F. Wysotzki, editors, *A Cognitive Science Approach to Reasoning, Learning and Discovery*, pages 55–82. Pabst Science Publishers, Lengerich, Germany, 1998. see also http://cogent.psyc.bbk.ac.uk/.

P. E. Fleiss and J. L. Shrout. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

C.R. Gallistel, Ann L. Brown, Susan Carey, Rochel Gelman, and Frank C. Keil. Lessons from animal learning for the study of cognitive development. In Susan Carey and Rochel Gelman, editors, *The Epigenesis of Mind*, pages 3–36. Lawrence Erlbaum, Hillsdale, NJ, 1991.

Steve Hanks, Martha E. Pollack, and Paul R. Cohen. Benchmarks, testbeds, controlled experimentation and the design of agent architectures. Technical Report 93–06–05, Department of Computer Science and Engineering, University of Washington, 1993.

Ian D. Horswill. *Specialization of Perceptual Processes*. PhD thesis, MIT, Department of EECS, Cambridge, MA, May 1993.

Ian D. Horswill. *Specialization of Perceptual Processes*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, May 1993.

N. Jakobi. Evolutionary robotics and the radical envelope of noise hypothesis. *Journal Of Adaptive Behaviour*, 6(2):325–368, 1997.

Hiroaki Kitano. Special issue: Robocup. *Applied Artificial Intelligence*, 12(2–3), 1998.

Hiroaki Kitano. Robocup rescue: A grand challenge for multiagent systems. In *The Fourth International Conference on MultiAgent Systems (ICMAS00)*, pages 5–12, Boston, 2000. IEEE Computer Society.

Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, and Eiichi Osawa. RoboCup: The robot world cup initiative. In *Proceedings of The First International Conference on Autonomous Agents*. The ACM Press, 1997.

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

R. S. Lockhart. *Introduction to Statistics and Data Analysis for the Behavioral Sciences*. Freeman, 1998.

R. M. Neal. Assessing relevance determination methods using DELVE. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, pages 97–129. Springer Verlag, 1998. See also `http://www.cs.utoronto.ca/~delve/`.

U. Nehmzow, M. Recce, and D. Bisset. Towards intelligent mobile robots - scientific methods in mobile robotics. Technical Report UMCS-97-9-1, University of Manchester Computer Science, 1997. Edited collection of papers, see also related special issue of *Journal of Robotics and Autonomous Systems*, in preperation.

David L. Parnas. Software aspects of strategic defense systems. *American Scientist*, 73(5):432–440, 1985. revised version of UVic Report No. DCS-47-IR.

Phoebe Sengers. Do the thing right: An architecture for action expression. In Katia P Sycara and Michael Wooldridge, editors, *Proceedings of the Second International Conference on Autonomous Agents*, pages 24–31. ACM Press, 1998.

Aaron Sloman and Brian Logan. Building cognitively rich agents using the Sim_agent toolkit. *Communications of the Association of Computing Machinery*, 42(3):71–77, March 1999.

L. A. Stein. Challenging the computational metaphor: Implications for how we think. *Cybernetics and Systems*, 30(6):473–507, 1999.

Kristinn R. Thórisson. *Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills*. PhD thesis, MIT Media Laboratory, September 1996.

Toby Tyrrell. *Computational Mechanisms for Action Selection*. PhD thesis, University of Edinburgh, 1993. Centre for Cognitive Science.