Increasing Generalization Requirements for Cognitive Models: Comparing Models of Open-ended Behavior in Dynamic Decision-Making

Christian Lebiere (cl@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15213 USA

Cleotilde Gonzalez (coty@cmu.edu) and Varun Dutt (varundutt@cmu.edu)
Dynamic Decision Making Laboratory, Department of Social and Decision Sciences, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15213 USA

Walter Warwick (wwarwick@alionscience.com)
Alion Science and Technology
4949 Pearl East Circle, Boulder, CO 80301

Keywords: Cognitive Architectures; Model Validation and Comparison; Dynamic Decision-Making.

Introduction

Model comparison is becoming an increasingly common method in computational cognitive modeling. The methodology is seemingly straightforward: model comparisons invite the independent development of distinct computational approaches to simulate human performance on a well-defined task. Typically, the benchmarks of the comparison are goodness-of-fit measures to human data that are calculated for the various models. Although the quantitative measures might suggest that model comparisons produce "winners," the real focus of model comparison is, or at least should be, on understanding in some detail how the different modeling "architectures" have been applied to the common task. And in this respect, the seemingly straightforward method of model comparison becomes more complicated.

The idea that a model comparison might be used to pick a winning approach resonates with common intuitions about model validation, namely, that a good fit is good evidence for the theory the model implements. But to the extent that model comparisons seek to illuminate general features of computational approaches to cognition rather than to validate a single theory of cognition, they depart from the familiar mode of good fit, good theory. Instead, a model comparison forces us to think about the science of modeling. A good fit is thus relegated to a necessary requirement rather than an end in itself, and the focus shifts toward a deeper understanding of the modeling approaches themselves. This shift brings into focus a host of new questions having to do with the relationship between model and architecture, theory and implementation, the relative contributions of the modeler and of the architecture to the final model, the role of parameter estimation in model development, the suitability of the simulated task to exercise features of the various architectures, the extensibility of the simulated task and the practical considerations that go into integrating disparate approaches within a common simulation environment. In this symposium, we address these issues in the context of our own model comparison. Our ultimate goal is to evolve a formal methodology to ensure the soundness of future comparison efforts and develop an infrastructure to make such efforts an ongoing process rather than one-off events.

Requirements

We have direct experience from a number of modeling comparisons projects, including the AFOSR AMBR modeling comparison (Gluck & Pew, 2005) and the NASA Human Error Modeling comparison (Foyle & Hooey, 2008). We have also entered cognitive models into multi-agent competitions (Billings, 2000; Erev et al, submitted) and organized symposia featuring competition between cognitive models as well as mixed human-model competitions (Lebiere & Bothell, 2004; Warwick, Allender, Strater and Yen, 2008). From these endeavors, we have gained an understanding of the required (and undesirable) characteristics of a task for such projects. While previous model comparison efforts did illustrate the capabilities of some modeling frameworks, the tasks were often ill suited to that purpose for a number of reasons:

- Some tasks demand a considerable effort just to model the details of task domain itself, which often results in a model whose match to the data primarily reflects the structure and idiosyncrasies of the task rather than the underlying cognitive mechanisms. This does not serve the primary purpose of a model comparison effort, which is to shed light upon the merits of the respective modeling frameworks rather than the cleverness and diligence of their users.
- Some tasks do not stretch model functionality beyond the conditions for which human data is available. The comparison effort can then be gamed by simply optimizing the model parameters to the data available, which puts frameworks that emphasize constrained, principled functionality at a disadvantage over those that permit arbitrary customization and parameterization.

- Likewise, some tasks are too specialized, emphasizing a single aspect, characteristic or mechanism of cognition and do not require the broad, integrated functional capabilities required of a general cognitive framework.
- If no common simulation or evaluation framework is provided, each team can focus on the aspects of the task most amenable to their framework, at the cost of making a direct comparison all but impossible.
- Finally, tasks for which no suitably comparable human data is available bias the effort toward a purely functional evaluation of model against model (rather than against data), which emphasizes performance at the expense of empirical fidelity.

This experience has taught us that the desirable characteristics of a task for a model comparison include:

- Lightweight, to limit overhead of integration, task analysis and knowledge engineering requirements.
- Fast, to allow efficient model development and collection of large numbers of Monte Carlo runs.
- Open-ended, to discourage over-parameterization and over-engineering of the model and test its generalization over a broad range of situations.
- Dynamic, to explore emergent behavior that is not predictable from the task specification.
- Simple, to engage basic cognitive mechanisms in a direct and fundamental way.
- Tractable, to encourage a direct connection between model and behavioral data.

Like other competitive benchmarks of human cognition (e.g. Robocup), the key is finding the right combination of simplicity, flexibility and emergent complexity.

Comparison Challenge

We believe the task we have selected, the Dynamic Stocks and Flows (Dutt & Gonzalez, 2007), meets these requirements and strikes the right combination between simplicity and complexity (Lebiere, Gonzalez, & Warwick, in press). The instructions to participate in this comparison challenge are on a web site[1], together with an executable version of the task, a text-based socket connection for models, and experimental data for a number of experimental conditions for model calibration. We collected data on additional conditions that were used to test the submitted model's generalization beyond the available conditions. Our focus in evaluating models was two-fold: quantitative measures of the models' fit to the data in the generalization conditions, and qualitative assessment of the generality and constraints of the underlying theories in meeting the demands of the task. The best entries under each criterion were invited to describe their model in this symposium.

Conclusion

A number of tests for a general theory of intelligence have been advanced (e.g. Cohen, 2005; Anderson & Lebiere, 2003). A key common aspect is to enforce generality in approach, in order to prevent special-purpose optimization to narrow tasks and force integration of capabilities. One can view that strategy as effectively overwhelming the degrees of freedom in the architecture with converging constraints in the data. However, precise computational specifications of those tests have to tread a tight rope between requiring unreasonable amounts of effort in modeling broad and complex tasks and falling back into narrow task specifications that will again favor engineered, optimized approaches. This model comparison challenge is our attempt at testing general cognitive capabilities in an open-ended manner by offering low barriers to entry in confronting different approaches with specific common problems that encourage integrated cognitive approaches.

[1] http://www.cmu.edu/ddmlab/modeldsf

Acknowledgments

This research was partially supported by the Advanced Decision Architectures of the Army Research Laboratory award number DAAD19-01-2-0009. Many thanks to Hau-yu Wong for her help in analyzing the data and to Mala Gosakan and Michael Matessa for their help in running the models.

References

Anderson, J. R. & Lebiere, C. (2003). The Newell test for a theory of cognition. *Behavioral & Brain Sciences 26*, 587-637.

Billings, D. (2000). The First International RoShamBo Programming Competition. *ICGA Journal*, Vol. 23, No. 1, pp. 42-50.

Cohen, P. (2005). If Not Turing's Test, Then What? *AI Magazine 26(4)*: 61–67.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S., Hau, R., Hertwig, R., Stewart, T., West, R., & Lebiere, C. (submitted). A choice prediction competition, for choices from experience and from description. *Journal of Behavioral Decision Making*.

Foyle, D. & Hooey, B. (2008). *Human Performance Modeling in Aviation*. Mahwah, NJ: Erlbaum.

Gluck, K, & Pew, R. (2005). *Modeling Human Behavior with Integrated Cognitive Architectures.* Mahwah, NJ: Erlbaum.

Dutt, V., & Gonzalez, C. (2007). Slope of inflow impacts dynamic decision making. Paper presented at the Conference of the System Dynamics Society.

Lebiere, C., & Bothell, D. (2004). Competitive Modeling Symposium: PokerBot World Series. In *Proceedings of the Sixth International Conference on Cognitive Modeling*, Pp. 32-32.

Lebiere, C., Gonzalez, C., & Warwick, W. (in press). Convergence and constraints revealed in a qualitative model comparison. *Journal of Cognitive Engineering and Decision Making.*

Warwick, W., Allender, L., Strater, L., & Yen, J. (2008). AMBR Redux: Another Take on Model Comparison. Symposium given at the *Seventeenth Conference on Behavior Representation and Simulation*. Providence, RI.