Journal of Artificial General Intelligence 2(2) 1-19, 2010 Submitted 2010-11-24
10.2478/v10229-011-0006-4

Editorial: Cognitive Architectures, Model Comparison and AGI

Christian Lebiere CL@CMU.EDU
Psychology Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213,USA

Cleotilde Gonzalez COTY@CMU.EDU
Dynamic Decision Making Laboratory
Social and Decision Sciences Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213,USA

Walter Warwick WWARWICK@ALIONSCIENCE.COM
Alion Science and Technology
4949 Pearl East Circle, Suite 200
Boulder, CO 80301, USA

Abstract

Cognitive Science and Artificial Intelligence share compatible goals of understanding and possibly generating broadly intelligent behavior. In order to determine if progress is made, it is essential to be able to evaluate the behavior of complex computational models, especially those built on general cognitive architectures, and compare it to benchmarks of intelligent behavior such as human performance. Significant methodological challenges arise, however, when trying to extend approaches used to compare model and human performance from tightly controlled laboratory tasks to complex tasks involving more open-ended behavior. This paper describes a model comparison challenge built around a dynamic control task, the Dynamic Stocks and Flows. We present and discuss distinct approaches to evaluating performance and comparing models. Lessons drawn from this challenge are discussed in light of the challenge of using cognitive architectures to achieve Artificial General Intelligence.

Keywords: Cognitive Architectures, Model Comparison, Dynamic Stocks and Flows

1. Introduction

At its creation over 50 years ago, the field of Artificial Intelligence (AI) was understood as having a dual goal, as articulated by Herbert Simon: "AI can have two purposes. One is to use the power of computers to augment human thinking. … The other is to use a computer's artificial intelligence to understand how humans think." (Stewart, 1994). Over time, rather than benefit

from a complementary relationship, these two goals have diverged, and the fields of AI and cognitive science have each matured as essentially separate disciplines. Artificial intelligence has become dedicated to the sole purpose of the creation of intelligent computer programs, irrespective of their relation to human cognitive processes. And, despite some initial success at tackling broad challenges, it has focused on increasingly narrow tasks, using equally specialized techniques. At the same time, cognitive science has taken the role of studying the processes of human cognition and has largely adopted the methods of cognitive psychology, dividing cognition into increasingly narrow fields and experimentally and computationally studying highly constrained and simplified laboratory tasks.

The Artificial General Intelligence (AGI) community has issued a call to revive the "Strong AI" goal of achieving integrated and general, rather than separate and specialized, intelligence. Although AI and cognitive science might one day be unified under this call, we still confront the considerable challenge of exploring an enormous design space of potentially intelligent systems and architectures. In this light, the divergence of AI and cognitive science is especially unfortunate given that human cognition may be the only example of general intelligence on offer. About 35 years ago, Allen Newell issued a challenge to go beyond the divide-and-conquer approach of cognitive psychology (Newell, 1973). His proposed solution was to develop integrated computational frameworks that would implement the invariant mechanisms of human cognition. Those Unified Theories of Cognition, realized as computational "cognitive architectures," could be used to model human cognition across its entire spectrum of application. As such, they provide a set of credible candidate architectures for achieving the AGI goal of general, integrated intelligence.

But even if the human cognitive architecture constrains the design space for an AGI system relative to all possible systems, understanding the scope and limits of different cognitive architectures as computational instantiations of general intelligence is far from trivial. The purpose of this special issue is to explore the merits of a comparative approach to understanding cognitive architectures as AGI systems. *Model comparison* is critical to achieving an integrated and general intelligence and for making scientific progress (Gluck, Bello, & Busemeyer, 2008). The goal here is not to advocate a single theory of cognition or to promote a particular computational architecture for AGI. Rather, the focus is on exploring a comparative methodology around which we might reconcile some of the now divergent aspects of AI and cognitive science in pursuit of AGI.

The structure and content of this special issue has been influenced by a modeling comparison challenge organized by the action editors of this special section (Lebiere, Gonzalez & Warwick, 2009; Warwick, 2009; Lebiere, Gonzalez, Dutt & Warwick, 2009). The comparison was based on the simulation of a generic dynamic decision making task, the Dynamic Stocks and Flows (DSF) (Dutt & Gonzalez, 2007; Gonzalez & Dutt, 2007). DSF was designed to be as simple and accessible as possible to computational modelers while focusing on two key ubiquitous components of general intelligence: the control of dynamical systems and the prediction of future events. A general call for participation was submitted to invite independent developers using distinct computational approaches to simulate human performance in DSF. Participants in this challenge developed computational models to simulate human performance on the DSF task in a variety of conditions. The goal was to reproduce human behavior, including learning, mistakes and limitations in such a way that their representations would generalize to *new* conditions of the task undisclosed to the modelers. Results from three of the nine models submitted were selected for presentation at the 2009 International Conference on Cognitive Modeling. Those models, the DSF simulation software, and the supporting data are all available on the comparison web site (http://www.cmu.edu/ddmlab/modeldsf).

This special issue presents our experiences and lessons learned through the model comparison challenge. In the remainder of the Editors' Introduction, we describe the DSF task including human performance in the laboratory as well as the simulation infrastructure we developed to support model comparison. Next, we describe some of the challenges we faced as organizers in understanding human performance in this task and drawing meaningful comparisons among models. We then describe how the difficulties in evaluating model performance on the DSF task point to more general issues in the quantitative comparison of human and model performance data. We then take a further step back and look at how a new understanding of quantitative comparison will ultimately support the pursuit of an AGI before concluding with a brief overview of the other contributions to this special issue.

2. Overview of the DSF Task and Human Performance in the Estimation and Comparison Phases

The stock management problem is that of controlling the level of an accumulating quantity by making decisions about levels of inflow and outflow. This is a generic problem, pervasive in everyday life, which arises at every temporal, spatial, and organizational levels (Cronin, Gonzalez & Sterman, 2009). For example, capabilities and competitive advantages arise from the accumulation of resources and knowledge (Dierickx & Cool, 1989; Sterman, 1989b); managers must control their cash flows to maintain adequate stocks of working capital; and production must be adjusted as sales vary to sustain sufficient inventory. The stock management problem has been investigated in many ways. For example, researchers investigated the perception of the building blocks of every dynamic system (Cronin & Gonzalez, 2007; Cronin, et al., 2009; Sterman & Booth Sweeney, 2002; Booth Sweeney & Sterman, 2000) where a dynamic system is reduced to its most essential elements: one *stock* (a resource that accumulates or depletes over time) and *flows* that alter the stock (an inflow that increases the stock or an outflow that decreases the stock). A conclusion from past years of investigation is that these simple stock problems are unintuitive and difficult, even in simple systems with a minimal number of variables, and even for highly educated people with strong technical backgrounds. In one experiment, for example, Booth Sweeney and Sterman (2000) presented highly educated graduate students at an elite university with a picture of a stock and graphs showing the inflow and outflow, then asked them to sketch the trajectory of the stock. Although the patterns were simple, fewer than half the participants were able to correctly sketch the path of the stock. This same effect has been reproduced in multiple experiments (Cronin & Gonzalez, 1997; Cronin et al., 2009).

2.1 The design and function of the DSF

The Dynamic Stocks and Flows is a simulation tool for studying learning and decision making in the context of simple stock management problems (Dutt & Gonzalez, 2007; Gonzalez & Dutt, 2007). Although the simulated stock problems are simple in the traditional sense (i.e., they have few elements to manage), the DSF is still dynamically complex (Sterman, 2000). The complexity arises from the interaction between decisions made and changes in the environment over time.

The DSF represents the essential elements of every dynamic system: a single *stock*, which represents an accumulation of discrete (e.g., units in inventory) or continuous (e.g., water) units; *inflows*, which increase the level of the stock; and *outflows*, which decrease the level of the stock. The goal of this task is to maintain the stock at a particular level or at least within an acceptable range. External inflow and outflow increase or decrease the level of stock, both of which are outside the control of decision makers. Stock levels are also influenced by the user's decisions of

inflow and outflow, which increase or decrease the level of the stock and are under the control of the user. Further, the level of the stock at time *t* depends upon the state of the system at the previous time *t-1*, a characteristic of dynamic systems called *interdependency* (Edwards, 1962). Also inherent in dynamic systems are feedback loops, where a variable can affect itself and other variables.

Figure 1 displays the graphical user interface for the DSF environment. The stock is represented graphically as a tank. In this version, the simulation represents continuous units of the stock as water in a tank. The markings on the left side of the tank represent the water level in the tank at any instant of time. There are 4 pipes connecting the tank, as shown in Figure 1. Two pipes labeled *User Inflow* and *Environment Inflow* are located on the input side and increase the level of stock in the tank; two pipes labeled *User Outflow* and *Environment Outflow* are located on the output side and decrease the level of stock in the tank.

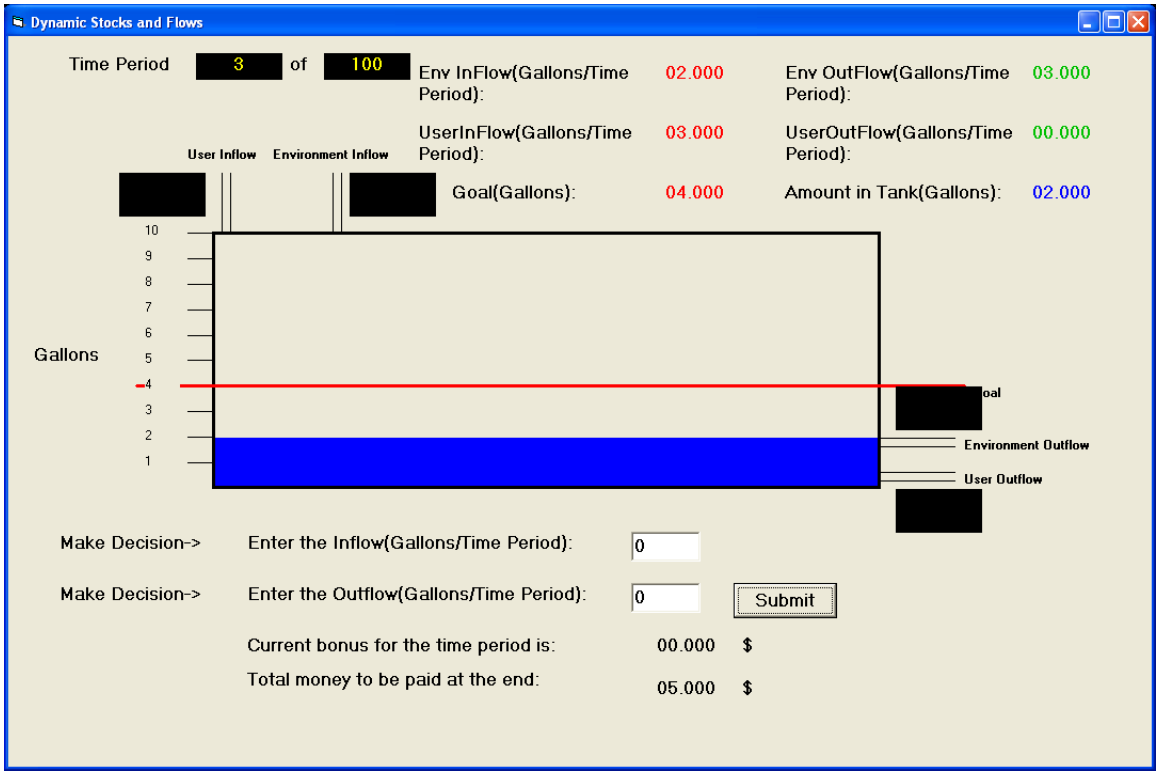


Figure 1. The interface of the DSF simulation. For a more detailed description of the different parts of the simulation, please see: Dutt & Gonzalez, 2007 and Gonzalez & Dutt 2007.

A user makes a decision on the inflow and outflow rates (*user inflow* or *user outflow*) by entering the values in the blank boxes at the bottom of the screen for each time period and hits *Submit*. The target level of accumulation is shown with a red horizontal line with *Goal* mentioned on the right side and also in the *Goal* information box. The current amount of water in the tank is shown in the *Amount in Tank* box. The Environment inflow and Environment outflow are exogenous functions that the user cannot control. After the user hits the submit button, the simulation determines the amount of water in the tank by adding the User Inflow and Environmental Inflow to the amount in the tank and subtracting the User Outflow and the Environmental Outflow. Then the simulation presents the resulting values in the following time

period. Thus, in the example of Figure 1, the goal is to keep the level of water at 4 gallons. At time period 2 a user entered 3 in the User Inflow box and 0 in the User Outflow box. The "Environment" added 2 as inflow and removed 3 as outflow. The resulting amount of water at time period 3 is 2 gallons.

2.2 The Estimation and Comparison Data Sets

The model comparison challenge was based on data from two experimental data sets, an estimation set and a comparison set. The two data sets were collected in the laboratory from human participants interacting with several different conditions of the DSF task. The estimation sessions were run and reported initially by Dutt & Gonzalez (2007). This estimation data set was provided to participants of the model comparison challenge to support the development of their models. They were given detailed data and were allowed to explore and study the estimation data set in relation to their own models. The goal given to them in the model comparison challenge was to develop a model that would predict human performance in the comparison data set.

2.2.1 THE ESTIMATION DATA SET

The estimation data set came from two experiments that varied with respect to the kind of function controlling the Environmental Inflow. In those experiments the Environmental Inflow was either an *increasing* function or a *decreasing* function over 100 trials. The Environmental Outflow function was constant and set to zero throughout 100 time periods. Hence, Environment net flow was equal to Environment Inflow. The main performance measure in DSF was the *goal discrepancy*: the difference between the goal and the stock amount in each time period. The subject's goal was to maintain the level of water in the tank within +/- 0.1 gallons from the 4 gallons goal during all 100 time periods. The initial water level in the tank was fixed in all conditions to 2 gallons.

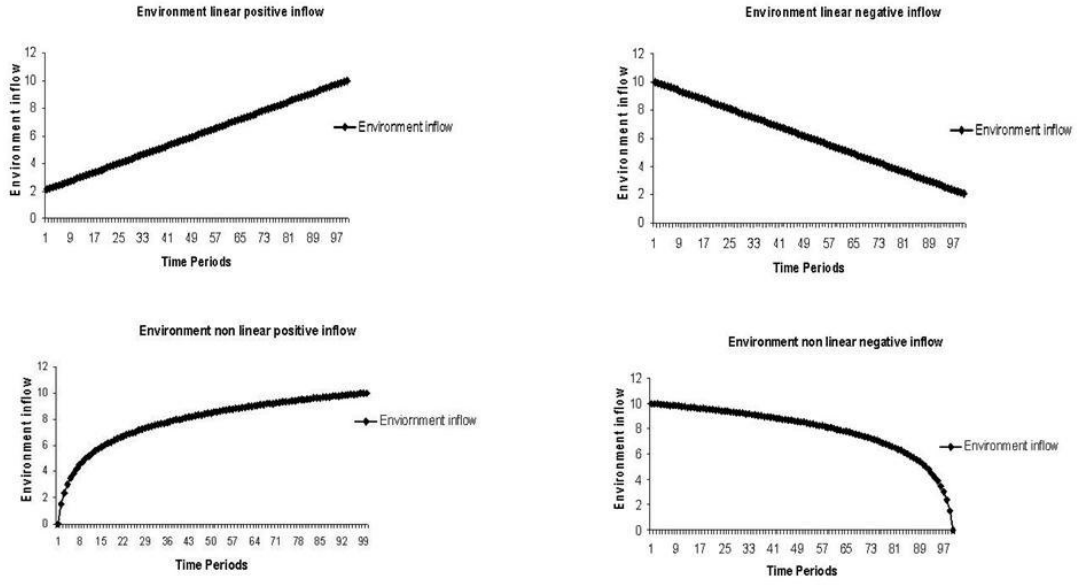


Figure 2. Functions of Environmental Inflow used in the estimation data set. Environmental Outflow was zero in all conditions.

In the first estimation data set, the Environment Inflow function was a Linear increasing or a Linear decreasing function (see Figure 2, top graphs). Environment Inflow increased or decreased over the course of 100 time periods using the formulas: 0.08 * (TimePeriod) + 2 for the increasing linear Environment Inflow function and -0.08 * (TimePeriod-1) + 10 for the decreasing linear Environment Inflow function. Both functions caused an equal amount of water (which was 604 gallons) to flow into the tank over the course of 100 time periods.

In the second estimation data set, the Environment Inflow function was a Non-Linear increasing or a Non-Linear decreasing function (see Figure 2, bottom graphs). Environment Inflow increased or decreased over the course of 100 time periods using the formulas: 5*LOG (TimePeriod) and 5*LOG (101- TimePeriod). Both conditions had a total Environment net flow of 831 gallons over the course of 100 time periods.

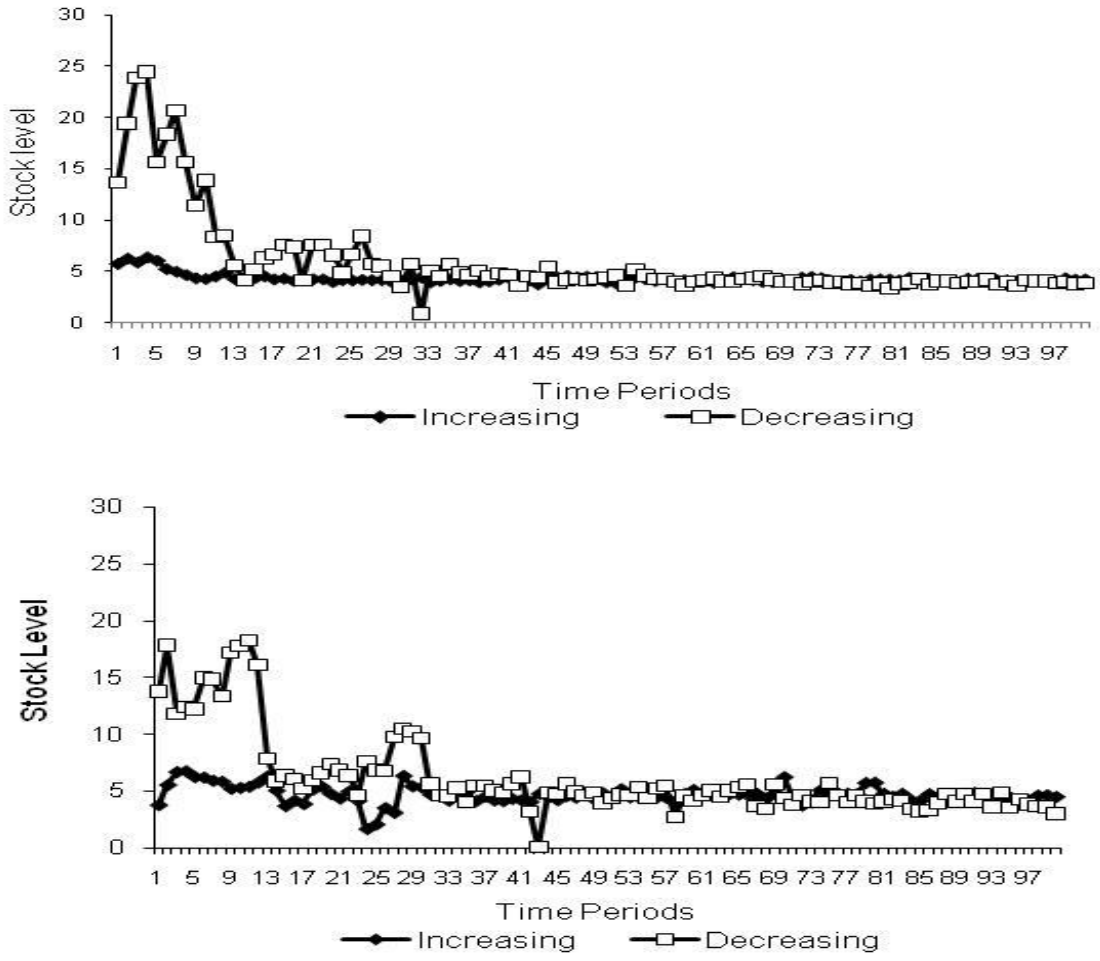


Figure 3. Human Performance results (stock level). Top Panel: results in the Linear increasing and decreasing functions. Bottom Panel: results in the Non-Linear increasing and decreasing functions.

The human performance results for the increasing and decreasing functions in DSF are presented in Figure 3. The raw data for each condition, included: participant id, time period

(from 1 to 100), Environmental Inflow value (according to the function in that condition) and Environmental Outflow value (zero in all conditions), User Inflow value and User Outflow value, actual amount in the tank, and target goal.

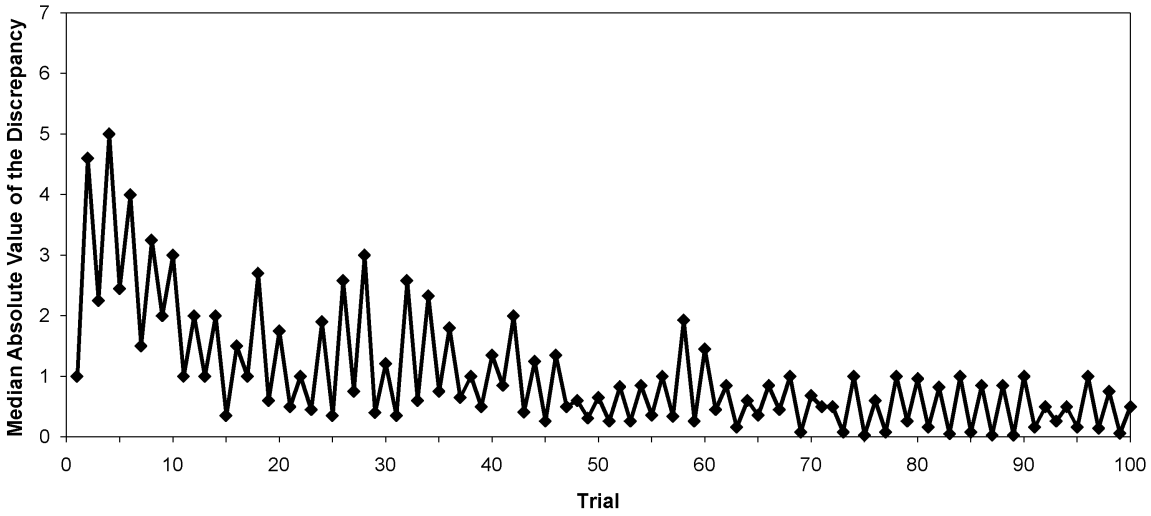
2.2.2 THE COMPARISON DATA SET

The comparison data were collected while the participants were working on their model development. The comparison data set consisted of five new conditions under two broad sets of manipulations, sequence and delay. There were a total of 120 participants in the laboratory study. Each participant in this data set (denoted by an ID number from 1 to 120) was tested in one condition under each of the two manipulations. The overall design was counterbalanced so that half the participants started with a sequence condition and the other half with a delay condition. In the sequence conditions the Environmental Inflow function generated a non-monotonic sequence of values of different length, including: a repeated sequence of length 2, the sequence of length 2 with noise, and a repeated sequence of length 4. The three conditions are as follows:

1) **Sequence=2**: The sequence of 1,5,1,5,... for 100 trials.
2) **Sequence=2+Noise**: The sequence 1,5,1,5… , +1 or -1 for each of 100 trials. The noise values were chosen randomly, thus the sequence after adding binary noise could be 0/2,4/6,0/2,4/6.... etc.
3) **Sequence=4**: The sequence of 0,4,2,6... for 100 trials.

This sequence manipulation was inspired by experiments in the field of sequence learning in cognitive psychology (e.g., Curran & Keele, 1993). The three separate conditions were tested with different human participants. All conditions started with 4 gallons of water in the tank, had a goal of 6 gallons, an Environmental Outflow of zero, and a total of 100 trials. Human participants received a base payment of $5 and bonus performance payment of 2.5 cents per trial, to make a total of $7.5 for 100 trials in approximately 30 minutes.

Figure 4 shows the median absolute value of the discrepancy Abs(Goal - Stock) over the 100 trials for the human data set in each of the three conditions above.



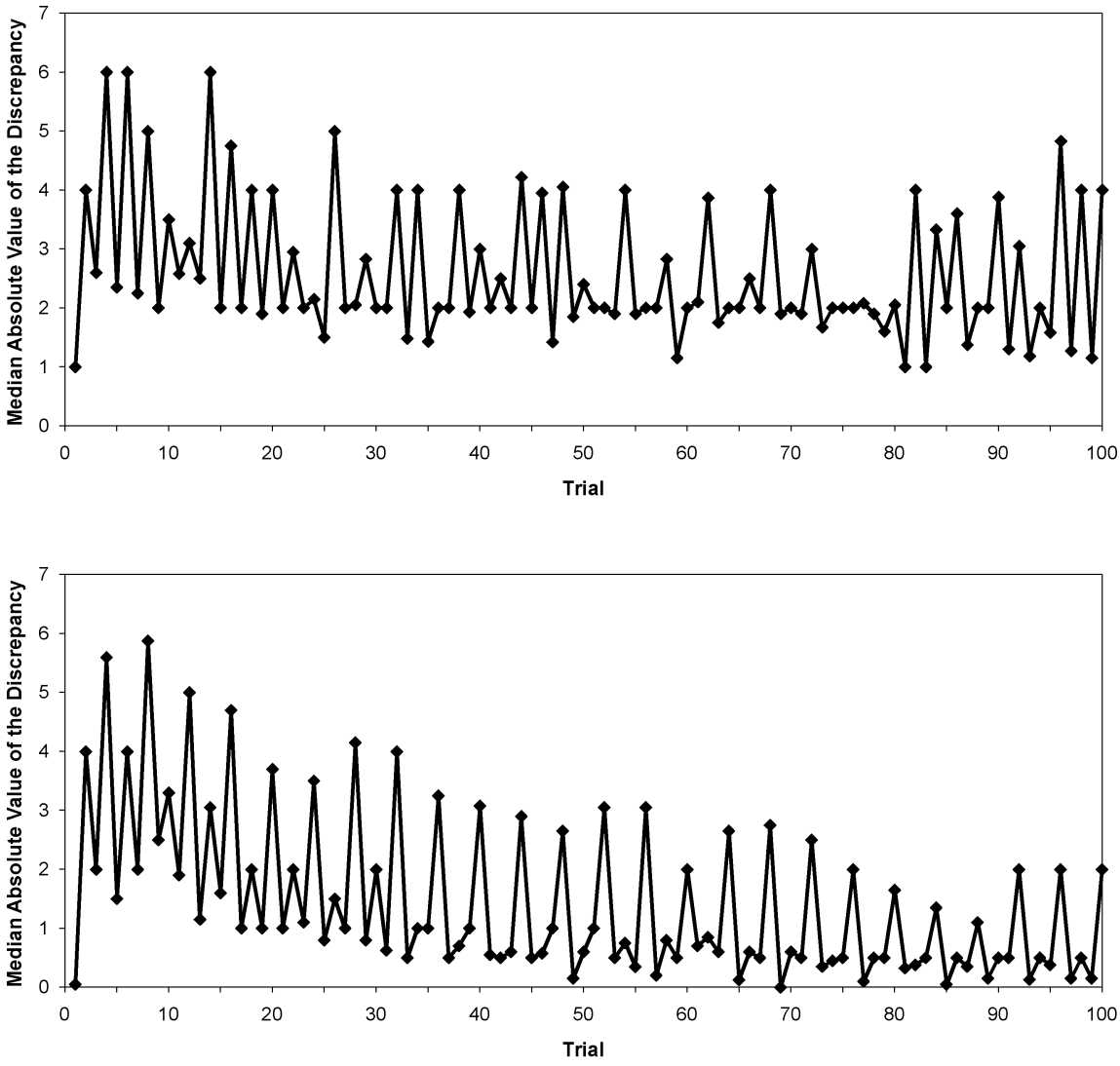


Figure 4. Median absolute value of discrepancy results for the Sequence conditions. Top Panel: Sequence 2. Middle Panel: Sequence 2 + noise condition. Bottom Panel: Sequence 4 condition.

In the delay conditions, user inputs to the system were delayed for either two or three time periods. This manipulation was inspired by the investigations of the detrimental effect of delayed feedback well known in the field of control systems (e.g., Brehmer, 1992). The Environmental Inflow function was the same linear increasing function used in the estimation data set (where there was an increase from 2 to 10 gallons of water deposited into the tank over the course of 100 trials). The two conditions are as follows:

1) **Delay=2**: All user inflow and outflow decisions were delayed for 2 time periods
2) **Delay=3**: All user inflow and outflow decisions were delayed for 3 time periods

Again, the conditions started with 4 gallons of water in the tank, had a goal of 6 gallons, an Environmental Outflow of zero, and a total of 100 trials. Human participants received a base payment of $5 and bonus performance payment of 2.5 cents per trial, to make a total of $7.5 for 100 trials in approximately 30 minutes.

Figure 5 shows the median absolute value of the discrepancy Abs(Goal - Stock) over the 100 trials for the human data set in each of the two conditions above.

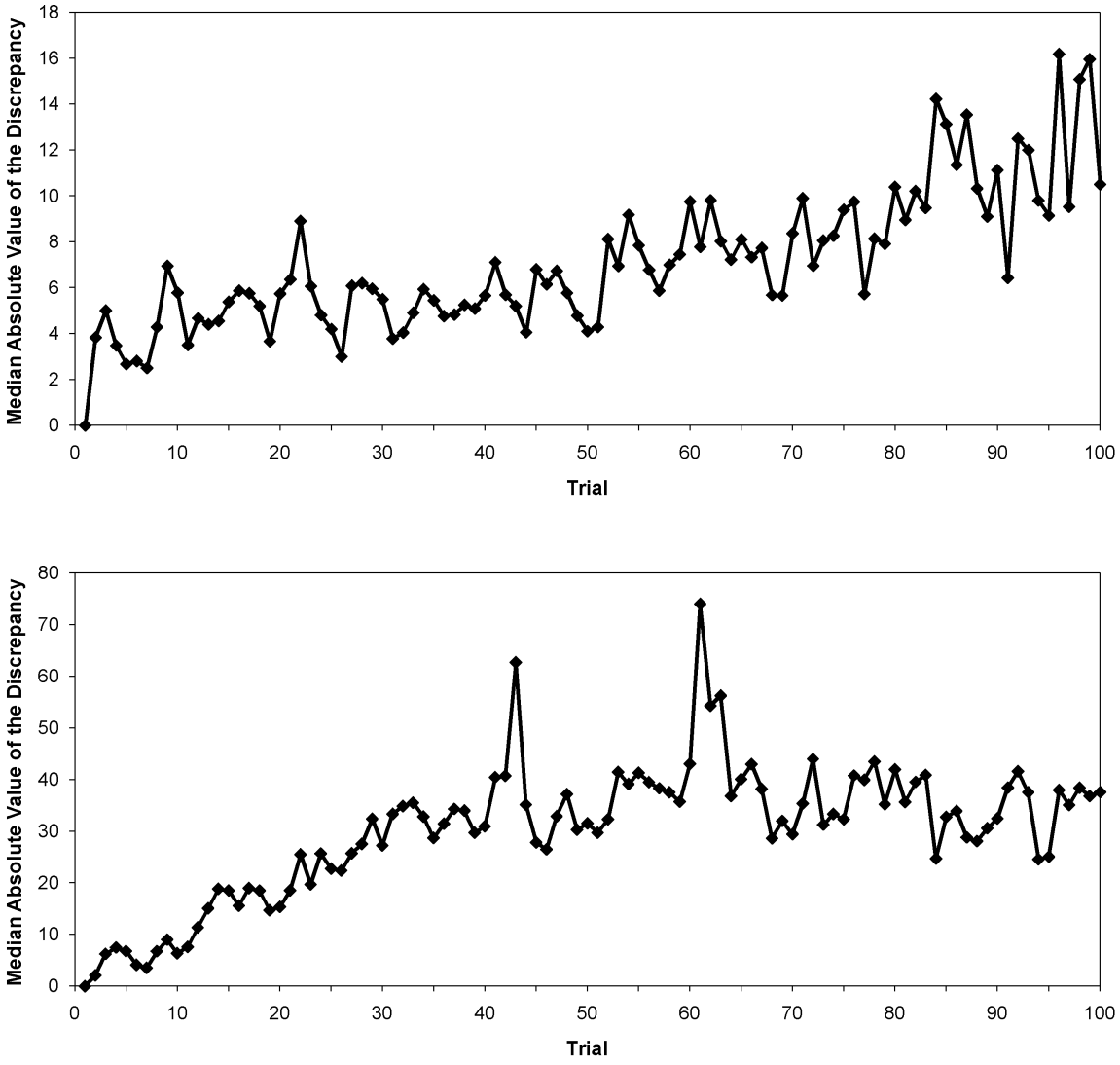


Figure 5. Median absolute value of discrepancy results for the Delay conditions. Top Panel: Delay 2 condition. Bottom Panel: Delay 3 condition.

2.3 Model Comparison Challenge Results

We received nine submissions to our modeling challenge. The models exhibited both a comforting set of common principles but also a fascinating diversity of approaches and emphases. We ran each of the models in all the conditions of the comparison data set. We faced the difficult task of performing a quantitative evaluation and ranking for purposes of adjudicating the challenge. We ran each of the models in all the conditions of the comparison data set. We computed Root Mean Square Error (RMSE) and correlation (R^2) measures over a range of potential values (bonus, user input and output, tank amount, discrepancy to goal) at the individual run level. Due to the strong correlation between the various values, we selected a single one, the discrepancy from the goal, as our focus. This left us with 10 quantitative measures, RMSE and R^2 over 5 distinct conditions. We rank-ordered each model over those 10 measures, with 1 being best and 9 worst, and summed those rankings over all 10 measures to establish an overall ranking. The ranks of each model were added, and accordingly the top three models were selected. The authors of the top three models were invited to present their work at the 2009 International Conference of Cognitive Modeling. These results are summarized in Table 1.

Figure 6 shows the median absolute value of the discrepancy over 100 trials for the three top models against the comparison human data in the sequence 2 condition, arguably the easiest of the transfer conditions. While the models show generally comparable average performance to the human data for the sequence 2 condition, they also display substantial differences, such as greater regularity and the gradual drift of the second model.

Figure 7 displays similar data for the Delay 3 condition, arguably the hardest condition for both models and subjects. This condition led to extremes values of discrepancy in the models in a way that made displaying the values difficult. For example, one of the models had values of absolute discrepancy as high as 40,000,000 gallons. In general, there was a tendency for the models to produce an erratic behavior with increased number of trials. Thus, to be able to plot the absolute discrepancy in a way that we could make some visual comparisons, we removed extreme values of absolute discrepancy that were above 2000 gallons. This is quite a high value of discrepancy, considering that the optimal value is zero. Although a couple of the human participants also produce extreme values of absolute discrepancy in some of the trials (again, on the order of 40,000,000 gallons), a large majority of trials from human participants stay below the 2000 gallon cutoff. Even after removing these extreme values, one can see in Figure 7 that the models, especially the first and second, display substantial swings in control (note the scale) beyond those of human subjects.

RSME												
Delay=2			Delay=3			Sequence=2			Sequence=2+Noise			Sequence=4		
Model	Value	Rank	Model	Value	Rank	Model	Value	Rank	Model	Value	Rank	Model	Value	Rank
4th Place Model	25559.65614	1	1st Place Model	5931065.843	1	6th Place Model	2.389391245	1	2nd Place Model	12.30789649	1	3rd Place Model	3.842132208	1
3rd Place Model	25560.64323	2	5th Place Model	5948339.426	2	1st Place Model	3.978684208	2	3rd Place Model	12.4796554	2	1st Place Model	4.120680522	2
5th Place Model	25560.73644	3	4th Place Model	5958227.153	3	3rd Place Model	4.330294541	3	1st Place Model	12.48050964	3	2nd Place Model	4.191526873	3
6th Place Model	25560.79757	4	3rd Place Model	5958290.296	4	2nd Place Model	4.446793766	4	5th Place Model	12.91020917	4	4th Place Model	4.324393357	4
1st Place Model	25560.84149	5	7th Place Model	5958302.643	5	5th Place Model	5.017039656	5	4th Place Model	13.20581246	5	5th Place Model	4.84942078	5
8th Place Model	25691.36149	6	8th Place Model	8623745.699	6	4th Place Model	5.333908033	6	7th Place Model	32.57013019	6	7th Place Model	12.22281907	6
2nd Place Model	104372.928	7	6th Place Model	10629371.62	7	7th Place Model	36.77878748	7	8th Place Model	39.28063253	7	8th Place Model	19.2543028	7
9th Place Model	913917.2672	8	2nd Place Model	15470739.89	8	8th Place Model	45.49902821	8	9th Place Model	165.6242666	8	9th Place Model	188.8278338	8
7th Place Model	1.74997E+37	9	9th Place Model	2.01691E+13	9	9th Place Model	149.2110442	9	6th Place Model	966873514.7	9	6th Place Model	1338101156	9

R^2												
Delay=2			Delay=3			Sequence=2			Sequence=2+Noise			Sequence=4		
Model	Value	Rank	Model	Value	Rank	Model	Value	Rank	Model	Value	Rank	Model	Value	Rank
7th Place Model	0.006945102	1	2nd Place Model	0.006305459	1	6th Place Model	0.032074972	1	9th Place Model	0.0046707 8	1	1st Place Model	0.048517273	1
2nd Place Model	0.001652099	2	1st Place Model	0.005542475	2	1st Place Model	0.023023154	2	2nd Place Model	0.004353249	2	4th Place Model	0.043084207	2
1st Place Model	0.000783014	3	4th Place Model	0.005026803	3	3rd Place Model	0.007556976	3	8th Place Model	0.00210721	3	3rd Place Model	0.040925091	3
5th Place Model	0.000766518	4	5th Place Model	0.003403337	4	4th Place Model	0.005576699	4	3rd Place Model	0.001874682	4	2nd Place Model	0.026398729	4
6th Place Model	0.000747421	5	6th Place Model	0.002676126	5	5th Place Model	0.004921518	5	1st Place Model	0.0010141	5	7th Place Model	0.009105476	5
4th Place Model	0.000621257	6	8th Place Model	0.000990819	6	7th Place Model	0.002348685	6	7th Place Model	0.000966799	6	6th Place Model	0.001432866	6
8th Place Model	0.000515284	7	9th Place Model	0.000717852	7	2nd Place Model	0.001906049	7	4th Place Model	0.000794785	7	5th Place Model	0.001201751	7
9th Place Model	9.61337E-05	8	7th Place Model	0.000625682	8	8th Place Model	5.8485E-05	8	6th Place Model	0.000304811	8	9th Place Model	8.96343E-05	8
3rd Place Model	1.23006E-06	9	3rd Place Model	3.12194E-05	9	9th Place Model	1.10954E-06	9	5th Place Model	6.29131E-05	9	8th Place Model	3.05454E-05	9

Table 1. Summary results of the participant models in the DSF Challenge.

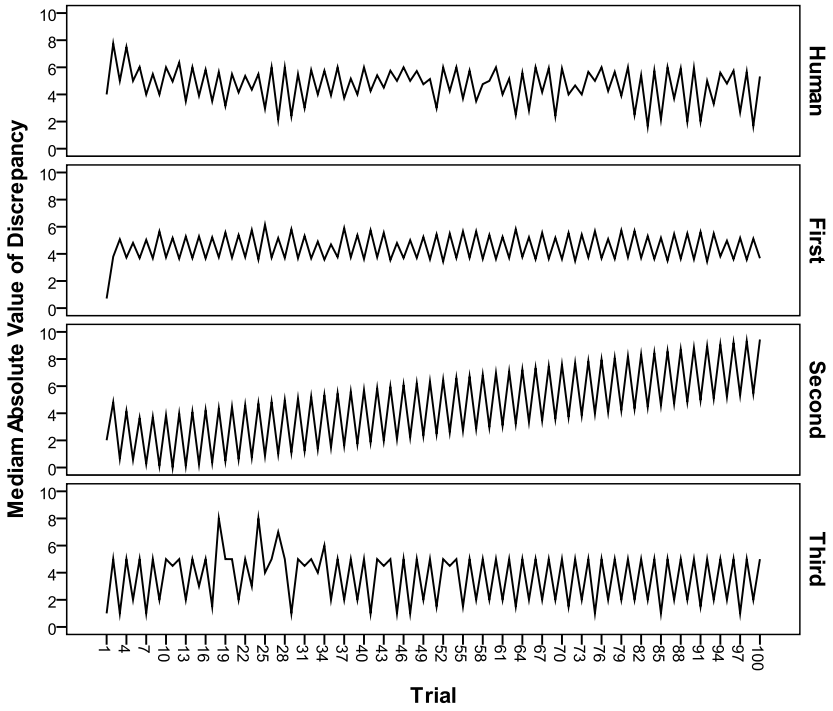


Figure 6. Median absolute value of discrepancy results for the comparison human and model data (first, second and third place) in the Sequence 2 condition.

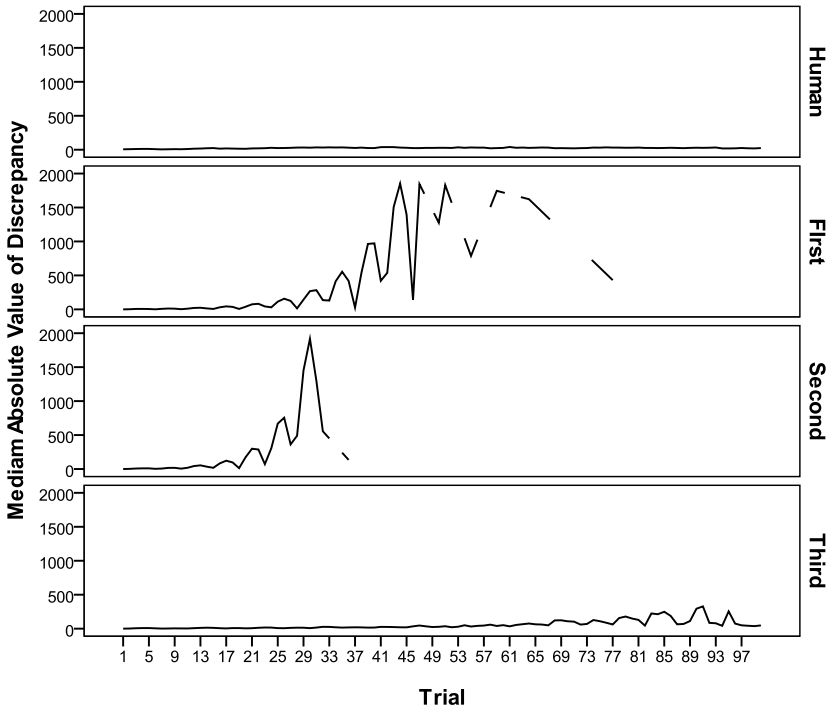


Figure 7. Median absolute value of discrepancy results for the comparison human and model data (first, second and third place) in the Delay 3 condition after removing extreme values.

No doubt there are serious issues with any quantitative measure in this case, including statistical validity of tests comparing similarly performing models under high-variability conditions. But an even deeper issue remains as to whether such traditional measures of fit would be appropriate to the DSF tasks even if the data were better behaved. (see Lebiere, Gonzalez & Warwick, 2009 for a discussion of qualitative and quantitative model comparison issues). Broadly speaking, the various models exhibited fairly similar strengths and weaknesses. As expected, they generally did better in the sequence conditions than in the delay conditions. The analyses below address a less traditional way in evaluating performance and performing comparisons between model and human data.

3. Domain-Specific Analysis

In evaluating performance in a complex task, including comparisons between model and human data, it is important to go beyond aggregate statistics that summarize overall performance. Instead, it is essential to develop quantitative measures that capture the key aspects underlying performance in the task and in particular those that reveal the fundamental cognitive mechanisms being used. The goal is to obtain the most direct measures that enable the modeler to understand and constrain the structure and content of the model instead of having to resort to post hoc and often ad hoc parameter optimization. These quantitative measures should be robust to superficial differences in performance, such as for example the particular phase of the oscillations in the sequence conditions performance, which could result in very poor or even negative correlation between models that might otherwise be quite similar, just because of some accident of timing. A key question is whether general quantitative performance measures can be found or whether they are bound to be specific to the particular domain and task. In the context of the DSF task this requires understanding the general cognitive functions that have to be applied to obtain effective performance. Those functions are basically twofold: controlling the simulation environment given its current state, and anticipating future environmental inputs that might affect future states.

We will apply our analysis here in the sequence+noise condition, which has a pair of interesting properties: (a) as explained below, the control function and the anticipation function result in inputs that are largely uncorrelated, which allows us to isolate those functions in the data, and (b) the noise limits the effectiveness of explicit reasoning strategies and the individual differences that result, allowing learning to proceed more smoothly over the entire length of the experiment.

The ability to control the system in its current state, assuming steady environmental inputs, reduces to a means-end focus on the discrepancy between current tank content and goal amount. Thus, the correlation between the discrepancy between current and desired tank amount at a given time step and the net user outflow should be a direct indicator of that ability. A perfect correlation would indicate a user's single-minded focus on eliminating the current discrepancy.

The ability to project the future state of the system and, in particular, to predict the major source of uncertainty (i.e., future environmental inputs) is harder to isolate. Specifically, we must separate the user's ability to predict future states from the ability to control the system, because users only express their knowledge of the system through explicit control. Here, we will measure the prediction ability by first taking the difference between the two quantities used to measure the control ability, namely, the discrepancy between current tank amount and the net user outflow, and then measure the correlation between that difference and the environmental inflow. Assuming the user's ability to control the system in its current state, a perfect correlation would

indicate that the user is also able to perfectly anticipate upcoming environmental flows to not only eliminate the current discrepancy (if any) but also the immediately upcoming one as well.

Figure 8 displays those quantities for each of the 40 individual subjects, with the first quantity along the abscissa and the second along the ordinate. The analysis confirms the substantial variation in performance in even this relatively simple task. The ability to control the system is generally quite good, with a correlation between current and desired amount discrepancy and user net outflow between 0.45 and 0.85 for most subjects. In contrast, the ability to predict the future system state, specifically the correlation between net environmental inflow and the difference between the user net outflow and the amount discrepancy, is quite low. It varies between 0.0 and 0.4 for most subjects. Thus, subject ability to anticipate environmental flows varies from non-existent to fairly substantial, especially considering the presence of random noise in the environmental inflow that limits its predictability. Finally, correlation between the two measures is low, with a negative relation between the ability to control the system and the ability to predict its future inputs (though some of that might be due to limitations of the approach that we have taken to separate the two measures).

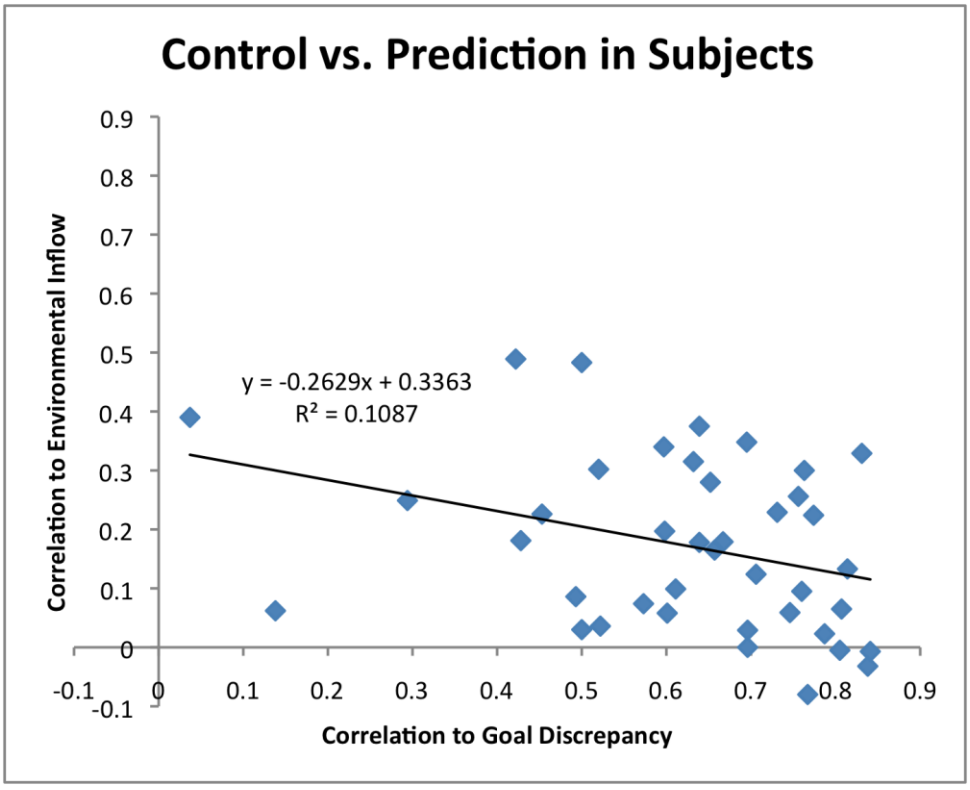


Figure 8. Individual Differences in Control and Prediction.

Figure 9 displays the same measures for each of 20 model runs for the top three models from our 2009 comparison challenge developed by Reitter (this issue), Iglesias et al. (Lebiere et al., 2009), and Halbruegge (this issue). The most striking aspect is the lack of variability in model performance. Reitter's model displays very strong control ability, with a correlation between net outflow control and discrepancy of over 0.9, but a strongly negative correlation between the

difference and future environmental inflow. This pattern results from the alternating nature of the environmental inflow sequence, which leads a model solely focused on reducing the current difference to systematically be one step behind the environmental input flows. Halbruegge's model displays more variation, especially in its prediction ability, but still a very high focus on control at the expense of a largely negative prediction ability. Iglesias' model is the closest to the average subject performance, with a control correlation of about 0.7 and a prediction correlation limited but at least positive around 0.1. However, that model is completely deterministic and fails to exhibit any of the variations of human performance.

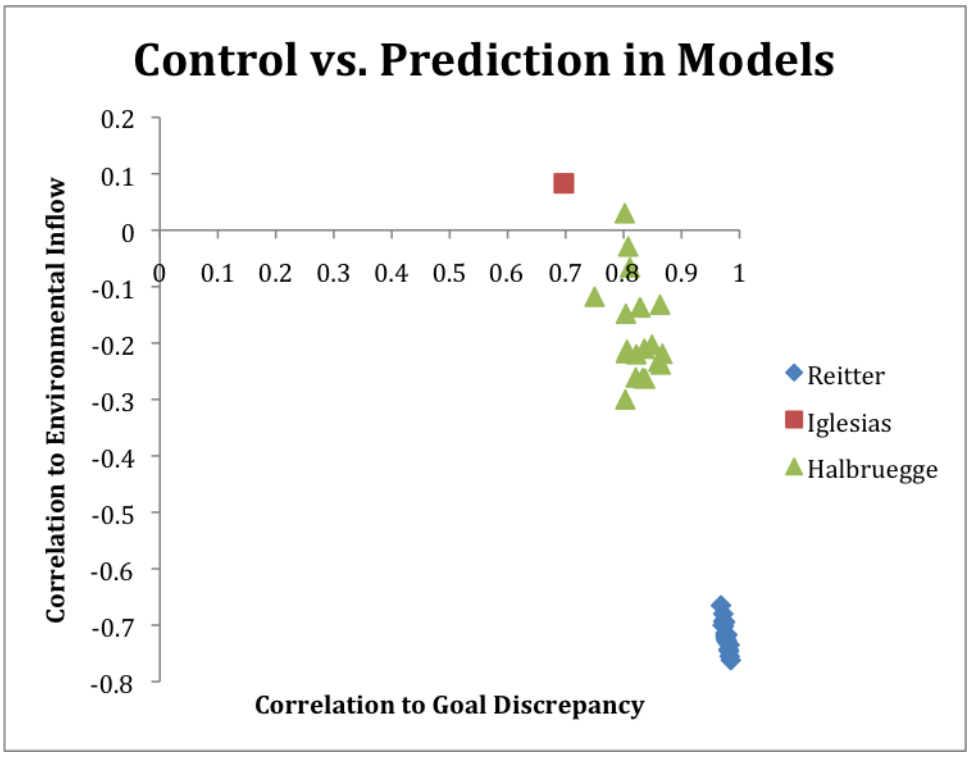


Figure 9. Model Performance in Control and Prediction.

Another important aspect of human behavior is the ability to improve performance with experience and adapt to changes in the environment. Just as individual differences reveal fundamental insights about the nature of cognition and its role in performance, learning and adaptivity provide stronger constraint on the possible nature of a model than average performance measures. In this analysis, we measure learning along the same dimensions, as the difference between the final 25 trials and the first 25 trials of the experiment. We chose to focus on the first and last quarters of the 100-trial runs because the former gives the best estimate of initial performance (a smaller window would be too noisy and a longer one would reflect too much learning already) and the latter of final performance (which tends to plateau after the three-quarter mark). Figure 10 indicates that most subjects get better at controlling the system between the first and last block of 25 trials. However, the ability to anticipate the next input again shows much more variation, with roughly equal numbers of increases or decreases across subjects. Part of this variation might reflect a meta-cognitive adaptivity: some subjects might be able to supplement

their baseline control ability with the additional ability to predict future environmental events, thus improving their overall performance, while others cannot achieve sufficient ability to effectively learn the environmental pattern, and might decide to stop trying to do so as it might detract from their first-order control ability when their guesses turn out erroneous. As for the averages in Figure 9, gains in the two measures appear only slightly correlated, with a negative relation between control gains and prediction gains, which might reflect the kind of metacognitive tradeoff just described.

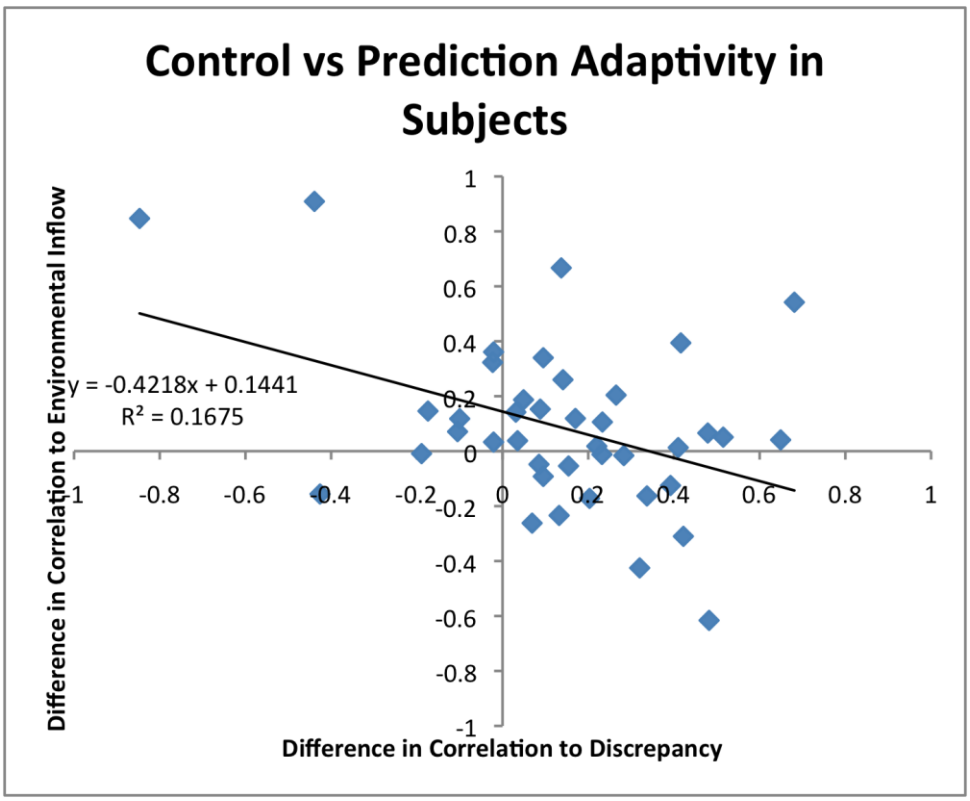


Figure 10. Subject Adaptivity in Control and Prediction.

Figure 11 shows that all models, like most subjects, become better at control over time. Reitter's model improves at control only slightly, largely because, as seen in Figure 9, that ability is almost already at ceiling to start with. However, it becomes slightly worse at prediction, possibly a result of the sole focus on control. Iglesias' model exhibits the same pattern but in a more extreme form, improving its control ability substantially but becoming significantly worse at prediction, both to a degree only matched by a pair of human subjects, again without variability. Halbruegge's model displays the pattern closest to human subjects, demonstrating an increase in control ability similar but somewhat more restricted than subjects (perhaps owing to a ceiling effect caused by its highest average) as well as a range of both positive and negative variation in control ability quite similar to the range of human subjects (albeit from a lower initial baseline). As for the average measures, Halbruegge's model also demonstrates the largest variation across individual runs.

The immediate conclusion of this analysis is that it is hard to determine a "best" model in a general way. Reitter's model fits best by the traditional measure of fit to overall performance used in the competition, Iglesias' model fits best by the average of the measures that we defined in our analysis, and Halbruegge's model fits best by the learning and adaptivity demonstrated in these measures as well as by the degree of variation in individual runs. This analysis established that the emphasis in model comparisons should be on understanding model performance in depth rather than on a competition aspect that uses a limited range of measures to declare a "winner".

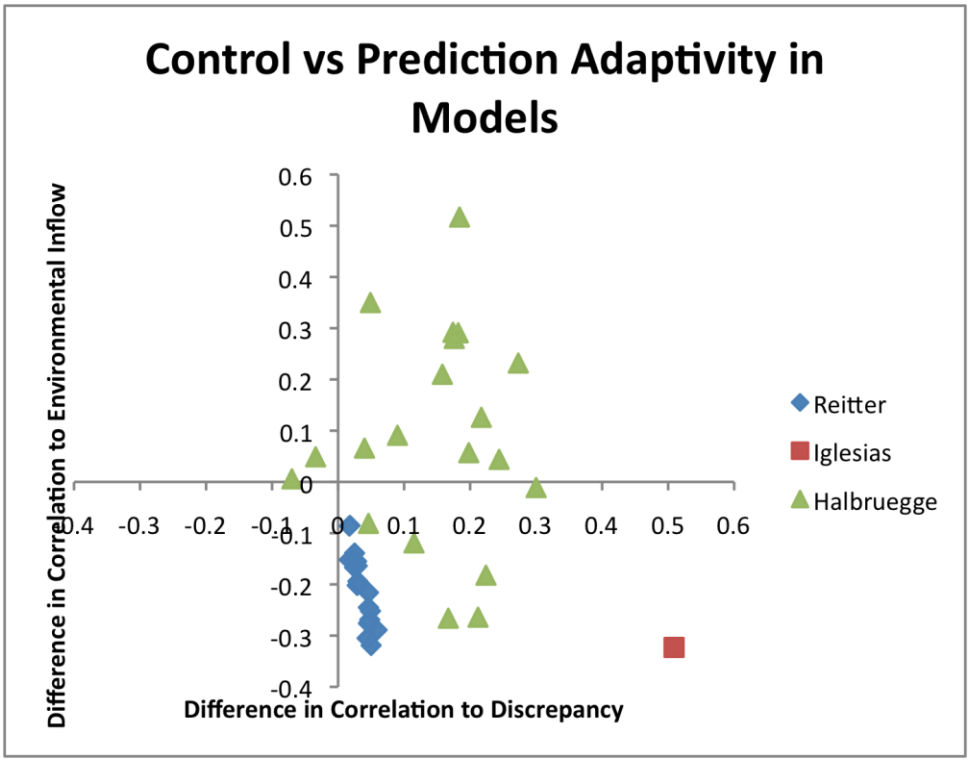


Figure 11. Model Adaptivity in Control and Prediction.

4. The Practice of Model Comparison and Progress toward AGI

It is easy to view a model comparison as a one-time event, but the scientific value of a comparison lies as much in the ongoing development and refinement of comparison techniques as in the identification of any particular winning model. Though implementing a predictive model of human performance for a task as demanding as the DSF is as instructive as it is difficult (as the contributions from Reitter, Halbruegge and Peebles & Banks demonstrate), we hope to advance the pursuit of AGI by pointing out the different contributions of our own model comparison.

First, despite the pointed advice of many of our colleagues, we still managed to underestimate some of the practical difficulties a comparison effort entails. In particular, it was difficult to communicate expectations to our participants and for some of them, in turn, to describe inner workings of their models to us. Clearly, a successful comparison requires that participants know and understand how their models will be connected to and interact with a simulated environment.

But this requirement is more than just a matter of ensuring software interoperability; it is also a matter of communicating what the models need to do, how we will verify that they have done what they were supposed to do and how it was that they managed to do what they did. Unfortunately, the vocabulary to communicate model capabilities and functions is limited and often the terms held in common among interdisciplinary research communities serve to confuse rather than clarify discussion. To the extent that the development of an AGI will likely depend on identifying and implementing general features of intelligence, it will be critical to understand and communicate exactly what those features are to a large and diverse community. Here, we repeat the advice of our colleagues: do not underestimate how hard it can be to communicate the content of a model and, by extension, the requirements and capabilities of an AGI.

Second, although well-defined tasks lead to well-behaved data and straightforward measures of fit, complex, dynamic tasks do not. Of course, it is the complex, dynamic tasks that make for interesting comparisons and are more likely to lie at the foundation of an AGI. As we discussed previously, the standard measures of fit are not always helpful in understanding model performance on such tasks. New measures and techniques are needed that will help us better understand when a model or an AGI is really doing its job. Stewart & West suggest such a measure while Gluck, Stanley, Moore, Reitter & Halbruegge describe how to judge whether a given component is really contributing to the overall performance of a model.

Finally, we think it is unlikely that fundamental progress toward AGI will be accomplished while implementing solutions to specific, one-off tasks. We chose the DSF task as our benchmark for comparison because it embodies general cognitive abilities like pattern detection and projection. Its generality ensured that it could be extended in ways that captured fundamentally different abilities within the same overall task definition, thus requiring no changes to models in order to test their generalization. Requiring models to generalize, even in seemingly obvious ways, is an important step away from over-fitted, data-specific models. Myers, Gluck, Gunzelmann & Krusmark march even further in this direction by describing how a model's performance can be validated against different time scales. The pursuit of an AGI is clearly furthered by model comparisons that require generalization in new and unpredictable ways. After all, perhaps the key feature of our intelligence is to adapt to unforeseen circumstances in effective, open-ended ways.

References

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

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Curran, T. & Keele, S.W. (1993). Attentional and nonattentional forms of sequence learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 189-202.

Dutt, V., & Gonzalez, C. (2007). Slope of inflow impacts dynamic decision making. In *Proceedings of the 25th International Conference of the System Dynamics Society* (pp. 79). Boston, MA: System Dynamics Society.

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.

Gonzalez, C., & Dutt, V. (2007). Learning to control a dynamic task: A system dynamics cognitive model of the slope effect. In Lewis, Polk, & Laird (Eds.), *8th International Conference on Cognitive Modeling* (pp. 61-66). Ann Arbor, MI.

Kaminka, G. A., & Burghart, C. R. (2007). Evaluating Architectures for Intelligence. *Technical Report WS-07-04.* AAAI Press, Menlo Park, California.

Lebiere, C., Gonzalez, C., Dutt, V. & Warwick, W. (2009). Increasing Generalization Requirements for Cognitive Models: Comparing Models of Open-ended Behavior in Dynamic Decision-Making. In *Proceedings of the 9th International Conference on Cognitive Modeling*. Manchester, England.

Lebiere, C., Gonzalez, C., & Warwick, W. (2009). A Comparative Approach to Understanding General Intelligence: Predicting Cognitive Performance in an Open-ended Dynamic Task. In *Proceedings of the Second Artificial General Intelligence Conference (AGI-09)*. Amsterdam-Paris: Atlantis Press.

Lebiere, C., & Wray, R. (2006). Between a Rock and a Hard Place: Cognitive Science Principles Meet AI-Hard Problems. *Technical Report SS-06-02.* AAAI Press, Menlo Park, California.

Newell, A. N. (1973). You can't play 20 questions with nature and win: Projective comments on papers in this symposium. In W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press.

Stewart, D. (1994). Interview with Herbert Simon. Omni Magazine, June 1994.

Warwick, W. (2009) Comparing the Comparisons. In *Proceedings of the Behavior Representation in Modeling and Simulation (BRIMS-09).* Sundance, Utah.