In D. Reitter & F. E. Ritter (Eds.), Proceedings of the 14th International Conference on Cognitive Modeling (ICCM 2016). University Park, PA: Penn State.

Cognitive Models of Prediction as Decision Aids

Christian Lebiere, Don Morrison (cl@cmu.edu, dfm2@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Tarek Abdelzaher, Shaohan Hu (zaher@illinois.edu, shu7@illinois.edu)
Department of Computer Science, University of Illinois at Urbana Champaign
201 N Goodwin Ave, Urbana, IL 61801 USA

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

Norbou Buchler,Vladislav D. Veksler (norbou.buchler.civ@mail.mil, vladislav.d.veksler.ctr@mail.mil)
Cognitive Sciences Branch, Human Resources and Effectiveness Directorate, U.S. Army Research Laboratory
Aberdeen Proving Ground, MD 21005 USA

Introduction

In the Age of Big Data, we are confronted with an increasingly rich and rapid flow of information. While the availability of data is increasing seemingly exponentially in our personal and professional lives, our basic human capabilities are not keeping pace. Recognizing with his customary foresight the increasingly deep disconnect between our abilities and the demands placed upon them, Herbert A. Simon once said, "Moore's Law fixes everything but us".[1]

Of course, technology has the potential to be the solution as well as the problem. Adaptive information retrieval tools such as search engines are helping us access and filter vast and diverse knowledge resources. In a more proactive way, personal electronic assistants such as Siri and Google Now offer to manage our data flows and provide us with timely, contextual information.

However, interpreting information and using it to make decisions is considerably more complex than simply making it available. Decision aids, including recommender systems, have been proposed to assist and delegate complex human decision-making. Leveraging the Big Data wave itself, those systems are typically data-driven, exploiting statistical regularities to extrapolate to similar situations. For instance, Netflix recently organized a competition to develop better algorithms for recommending movies by relying on ratings of viewers with similar tastes to a given customer. A fundamental problem with this approach is its opaque nature. When it fails, it tends to do so in ways unexpected and incomprehensible to human users, undermining trust in a system that not only performs poorly but cannot explain its own failures.

One potential solution is to design personal assistants that work in ways similar to humans, making them both more transparent and more compatible. Recently, a number of proposals have been made to measure artificial intelligence in more effective ways than the classic Turing Test, in particular by having it perform more typical tasks in human-like ways (AI Magazine, 2016). Going even further, the suggestion has been made to design intelligent agents based on the structure and mechanisms of the human brain (e.g., Stocco et al., 2010). For purposes of decision aids, such biologically inspired cognitive architectures might be a bridge too far. For instance, Google's PageRank algorithm might work in a way roughly similar to human associative memory, but few users would presumably care whether it mimics the structure of the hippocampus or the posterior cortex.

Cognitive architectures and models have primarily been developed as computational instantiations of theories of cognition. For purposes of serving as decision aids to human users, it is tempting to adopt the traditional AI view of treating them as black boxes and arguing that compatibility

[1] While Moore's Law might finally be running out of steam, it has been replaced by the exponentially increasing availability of massively distributed computation and sensor resources (Kurzweil, 2006).

with the human decision maker is primarily required from a functional, behavioral point of view. In terms of the Marr levels of analysis (Marr & Poggio, 1976), that would mean that what matters is primarily their functionality at the computational level rather than the algorithmic level or the implementational level. We disagree with that view.

Instead, we argue that, while the implementational level might not be directly relevant other than perhaps for purposes of scalability and efficiency, compatibility with the human decision maker at the algorithmic level is essential for truly effective interaction. Computational equivalence only enables a relatively superficial integration of outcomes, while algorithmic equivalence enables a deeper integration of processes.

We illustrate the distinction by introducing two functions of a cognitive model as decision aid: prediction and source selection. Prediction involves generating a recommended decision for the user to follow, and as such only requires computational compatibility. However, it leaves the user with little choice beyond accepting or rejecting the recommendation in its entirety. Source selection consists in selecting a subset of information on which the human user would base his own decision. While this enables richer interaction between user and decision aid, it also requires deeper compatibility, down to the algorithmic level, because the selection process requires integration with the processes of the human decision maker.

In the rest of this paper, we introduce a decision-making task based on a real world data set of emergency situations. We then describe a model of the task based on a rational analysis of cognition, and present quantitative results. Finally, we discuss implications for the design of cognitively inspired decision aids and recommender systems, and point out future work directions.

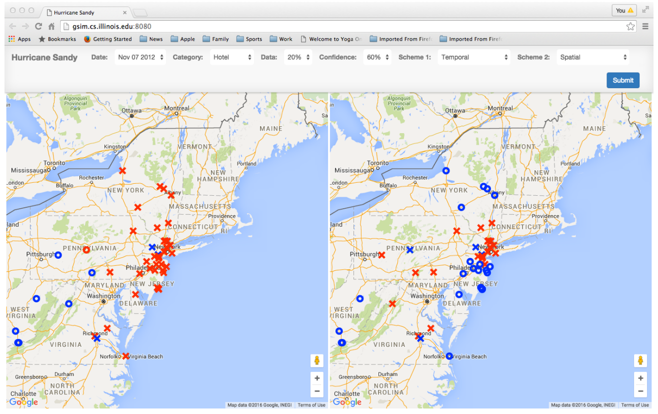
Task and Data

We focus on the problem of data extrapolation in participatory sensing applications, where users both use and provide information to the system, in the face of disruptive pattern changes, such as those that occur during natural disasters. We consider cases where resource limitations or accessibility constraints prevent attainment of full real-time coverage of the measured data space, hence calling for data extrapolation. Many time-series data extrapolation approaches are based on the assumption that past trends are predictive of future values. These approaches do not do well when disruptive changes occur. An alternative recourse is to consider only spatial correlations. For example, certain city streets tend to get flooded together after heavy rain (e.g., because they are at the same low elevation), and certain blocks tend to run out of power together after a thunderstorm (e.g., because they share the same power lines). Understanding such correlations can thus help infer state at some locations from state at others when disruptive changes (such as a flood or a power outage) occur.

We evaluate our prediction model through a real-world disaster response application. In November 2012, Hurricane Sandy made landfall in New York City. It was the second-costliest hurricane in United States history (surpassed only by hurricane Katrina) and the deadliest in 2012. The hurricane caused widespread shortage of gas, food, and medical supplies as gas stations, pharmacies and (grocery) retail shops were forced to close. The shortage lasted about a month. Recovery efforts were interrupted by subsequent events, hence triggering alternating relapse and recovery patterns. The daily availability of gas, food, and medical supplies was documented by the All Hazard Consortium (AHC), which is a state-sanctioned non-profit organization focused on homeland security, emergency management, and business continuity issues in the mid-Atlantic and northeast regions of the United States. Data traces[2] were collected in order to help identify locations of fuel, food, hotels and pharmacies that may be open in specific geographic areas to support government and/or private sector planning and response activities. The data covered states including WV, VA, PA, NY, NJ, MD, and DC. The information was updated daily (i.e., one observation per day for each gas station, pharmacy, or grocery shop).

With these points of interest sites and input data as ground truth, we evaluate the model predictions. The metrics we use are accuracy of inference and amount of data needed. We break time into daily cycles to coincide with the AHC trace. We then plot the performance of the model when a configurable amount of today's data is available (in addition to all historic data since the beginning of the hurricane).

We evaluate the solutions on November 3rd, and November 8th. November 8th corresponds to a period of disruptive change due to a second snowstorm that hit after Sandy, causing massive temporary relapse of recovery efforts due to new power outages, followed by a quick state restoration to the previous recovery profile. November 3rd is an example of a period of little change, when damage was incurred but recovery efforts have not yet been effective. The same trend was observed for all datasets, namely, gas, pharmacy, and food.



Hurricane Sandy

gsim.cs.illinois.edu:8080

Hurricane Sandy Date: Nov 07 2012 Category: Hotel Data: 20% Confidence: 60% Scheme 1: Temporal Scheme 2: Spatial

Figure 1: Data Extrapolation Task Interface.

[2] Available at: http://www.ahcusa.org/hurricane-Sandy-assistance.htm

Figure 1 displays a snapshot of the interface that we used to display the model results during model development. While we will focus in this paper on quantitative results for model evaluation, visualizing the spatial and temporal patterns of model prediction helped us understand the workings of the model and its strengths and shortcomings. It also helped us experiment with the model efficiently in exploring parameter settings and comparing model versions. Pull-down menus let the modeler easily select the date of the comparison, the category of data (pharmacy, food, gas), the sampling rate (percentage of the day's data to use in addition to historical data), the confidence threshold (probability to label an outlet as open, defaulting at an unbiased 50%) and any two versions of the model (see next section) to compare side by side against each other. Circles/crosses indicated a prediction that the outlet was open/closed. Color indicated the correctness of the prediction, with blue and red representing correct and incorrect respectively, while grey indicates no prediction was made because that data point was sampled. The data points were plotted on a Google Maps overlay of the geographical area, allowing the modeler to zoom in and out on various areas.

Cognitive Model

While prediction can be viewed as a specialized exercise best left to domain experts and statisticians, e.g., weather forecasting, stock market investing, sports betting, it also forms the implicit basis of many common everyday tasks. Previous models have shown its ubiquity in domains ranging from game playing (West & Lebiere, 2001) to sports (Lebiere et al., 2003), decision-making (Erev et al., 2010), and learning event sequences (Wallach & Lebiere, 2000).

While prediction can require the use of elaborate strategies and expert knowledge, those approaches are highly domain-specific and thus generalize poorly and tell us little about the basic nature of cognition. More fundamentally, complex approaches still seem to rely on a common basis of implicit statistical inference (e.g., Oaksford & Chater, 2007). The rational analysis of cognition (Anderson, 1990) has argued that our cognitive mechanisms have evolved to reflect the statistical structure of the environment. These regularities are quite pervasive and are displayed by our cognitive systems even when they are unwarranted and result in cognitive biases (Lebiere et al.You, 2013).

The rational analysis of cognition can offer a computational-level account of cognitive prediction. To achieve an algorithmic account with constrained quantitative predictions, we used the ACT-R cognitive architecture (Anderson et al., 2004). The mechanisms of its declarative module, in particular, reflect pervasive statistical patterns of the environment such as the power laws of learning and forgetting. As prediction relies on the knowledge of past events, it is logical to base the model on retrieval of information from long-term declarative memory.

In ACT-R, information is represented in declarative memory in the form of chunks, which are structured objects consisting of a set of attributes (also known as slots) with associated values. Chunk complexity (i.e., number of attributes) is typically limited, reflecting capacity constraints such as the size of working memory (Miller, 1956; Cowan, 2001). For instance, it would be unreasonable to store the entire history of an outlet or a whole day's data in a single chunk. Beyond capacity limitations, theories of chunk creation also typically limit their content to information that is available simultaneously at a given point in time and thus can plausibly be bound together in a new chunk structure.

Therefore, each chunk in memory represents the availability of a given outlet on a given day. Attributes that are represented include the identity of the outlet (itself represented as another chunk) and its status: open or closed (also represented as a chunk). The specific day could have been represented as a third attribute, although we decided against it for two reasons. First, it is slightly implausible that people would explicitly label each memory with the date of the day in which it was formed. Second, it would have resulted in a proliferation of memory chunks (one for each day and outlet) without markedly affecting the model predictions when the blending mechanism is used (see below).

Instead, time is represented implicitly in the activation of the corresponding chunk. The base-level activation B_i of a chunk i reflects its history of (re)creation and access as follows:

$$B_i = log \sum_{j=1}^{n} t_j^{-d} \qquad (1)$$

Where t_j is the time lag since the *j*th occurrence of the chunk, n is the total number of occurrences, and d is the decay rate (typically fixed at 0.5, as is the case in this model). For any given outlet, at most two associated chunks exist in memory: one recording that the outlet is closed and another recording that it is open on a given day. The base-level activation of these chunks will be reinforced with each occurrence of the respective event. The temporal version of the model then obtains a prediction for the status of a given outlet by retrieving the most active chunk associated with that outlet and returning the status stored in that chunk. Because the total activation A_i of chunk i also includes in addition to the base-level activation a stochastic component controlled by noise parameter s (using the typical value of 0.25 here), the retrieval process is probabilistic, described by the probability *P(i)* following the Boltzmann (softmax) distribution over all candidate chunks *j*:

$$P(i) = \frac{e^{\frac{A_i}{s}}}{\sum_j e^{\frac{A_j}{s}}} \qquad (2)$$

For a given outlet, only two chunks will compete for retrieval, and the winning chunk will reflect a combination of frequency and recency of the associated outcome, which is generally the temporal properties that are desired for prediction.

However, as mentioned earlier, temporal criteria are of limited usefulness when facing sudden disruption such as natural disasters. While a given outlet is usually open (frequency), and was open yesterday (recency), it may not be open today if a disaster event happened in the meantime. In that case, spatial factors constitute an additional basis for making predictions. Assuming lack of specific event knowledge, e.g., where the storm happened to hit, the most direct basis for including spatial factors is the limited known availability of nearby outlets. In the absence of additional semantic information (e.g., the outlet brand), the most direct information to use when attempting to generalize across outlets is their spatial location.

Specifically, the spatial component of the model makes use of the partial matching mechanism in memory retrieval, which allows for chunks that do not exactly match the requested pattern to be considered for retrieval, but with a penalty that reflects the degree of mismatch. Specifically, the activation A_i of chunk *i* is now the sum of the base-level activation and a mismatch penalty term:

$$A_i = B_i + MP * \sum_k Sim(v, d) \qquad (3)$$

where *MP* is a mismatch penalty scaling parameter (set in this model at a fairly standard value of 2.0) applied over all *k* pattern components specified in the retrieval request (only the outlet identity in this case) and *Sim(v,d)* is the similarity penalty between the corresponding value *d* requested and the actual value *v* present in the chunk. To avoid introducing needless free parameters, the similarity between outlet chunks is set to a linear function of the geographic distance between them, scaled such that a distance of 25 miles corresponds to a penalty of 1 unit of activation.

When making a prediction for a given outlet, the model will therefore not only consider the history of that given outlet as expressed in the base-level activation of the two associated chunks, but also chunks associated with other outlets as well, with a preference for those closer to the given outlet. Note that unlike that is the target of the prediction, some of those outlets will a known status for the present day, significantly increasing the base-level activation of the corresponding day. Thus the retrieval process will reflect a competition between the recency (and frequency) of outcomes, as reflected in the base-level activation, and its (spatial) relevance, as reflected in the mismatch penalty term.

The final component of the model concerns how to aggregate the relevant knowledge. As specified in the retrieval equation (2), one could simply select the most relevant chunk and return the associated outcome (open or closed). However, that would leave the prediction relying on a comparatively small piece of information, e.g., a chunk of limited relevance being retrieved purely through recency bias or simply the stochasticity of the process. To reflect people's ability to weigh a sizable part of their knowledge base when making predictions (e.g, Lebiere, 1999), the blending retrieval mechanism specifies how to return a value *V* (in this case, the availability prediction of a specific outlet) that reflects the consensus of the entire set of considered chunks, weighted by their respective probability of retrieval *P(i)*:

$$V = argmin \sum_i P(i) * (Sim(V, V_i))^2 \ (4)$$

Where *Sim(V, V_i)* is the similarity between the consensus value *V* and the value V_i proposed by chunk *i.* In this model, those values returned by the retrieval process are the outlet availability values: open or closed. Treating those values as binary would result in a process where the evidence for each outcome in the form of the activation of the chunks representing that outcome would be weighted against that of the competing outcome, and the greater one selected.

However, a more general decision process is also possible. By setting those values as numerical outcomes (e.g., 1 for open and 0 for closed) and assuming linear similarities in that range (the default, as for distance similarities earlier), the consensus value *V* will be somewhere in that interval reflecting the degree of preponderance of one outcome over the other. That value can then be interpreted as a confidence value in the open outcome, and assessed against a probability decision threshold (as mentioned in the description of Figure 1). This reflects the requirements of real world applications, e.g., where one might not want to predict that an outlet is open during an emergency without a fairly high certainty. However, we will only consider majority decisions (i.e., probability threshold of 0.5) in the following results section.

Results

In the absence of comparable human data, we examine the prediction performance of the model on a functional basis, but also looking to assess its cognitive plausibility. We also report results for the temporal and spatial versions of the model to assess the relative contribution of the two mechanisms.

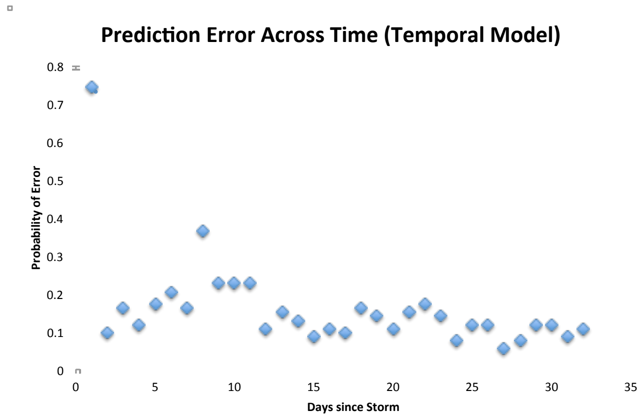


Figure 2: Performance of Temporal Model Across Time.

Figure 2 reports the aggregate performance of the temporal model across the entire range of data for about a month after the storm. This is the version of the model that only matches chunks for that specific outlet and relies only on its history. Performance is very poor on the day following the storm because of the lack of relevant data, but improves very quickly, with even a single day worth of data, because of the importance of the recency factor. Performance actually regresses slightly after that, as outlets become available again in a pattern that is difficult to predict, especially without access to semantic data such as outlet brands, which might get resupplied at the same time.

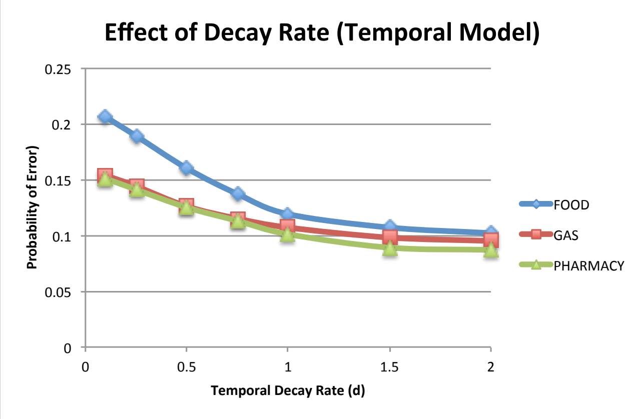


Figure 3: Effect of Decay Rate on Temporal Model.

Performance especially degrades on day 8, following a secondary storm that disrupts the pattern again. After that, it gradually improves over time to about 10% errors. Following the strong suggestion of the importance of the recency effect, Figure 3 examines the performance of the temporal rate as a function of the power law decay rate *d* for each outlet category averaged over all days, separated by outlet category. In general, a higher decay rate results in a lower error rate, indicating the primacy of recency over frequency. The availability of food outlets tends to be harder to predict than gas or pharmacy outlets, perhaps because their merchandise is more important or more perishable, leading to faster depletion, but the pattern is similar.

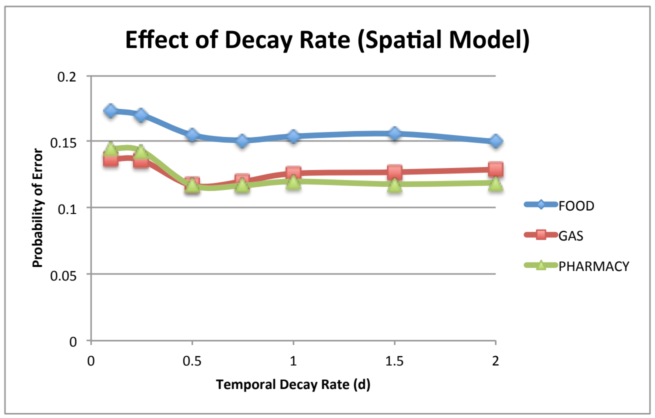


Figure 4: Effect of Decay Rate on Spatial Model.

Figure 4 reports the effect of the same decay rate, but for the spatial[3] model, that also reflects generalization across outlets using the partial matching and blending mechanisms. One can see that there is now a penalty for very high decay values that overemphasize recency. When considering a broader knowledge base, frequency of occurrence becomes more important and balances out against recency around the decay rate value of 0.5 that has become the standard value in ACT-R models for capturing human performance.

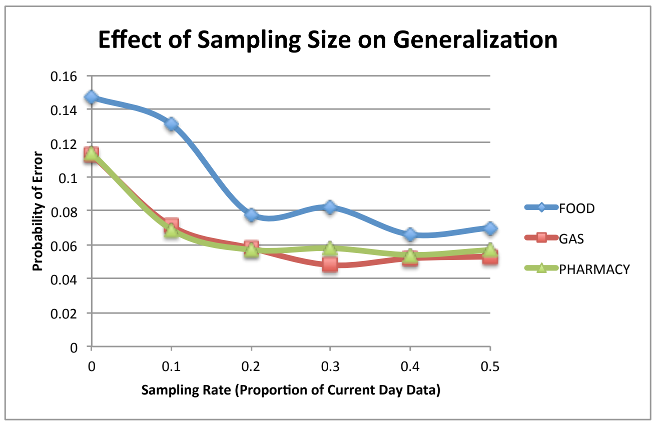


Figure 5: Effect of Sampling Size on Spatial Model.

The results of the spatial model presented in Figure 4 are actually slightly worse than those of the temporal model because we evaluated them on common ground, i.e., without including any of the current day's data for the spatial model to generalize from. Figure 5 examines the impact of the sampling rate of data for the current day to determine the effectiveness of the spatial model to generalize from nearby outlets. Generalization is quite effective, reducing the probability of error by half with about 20% of the current data. Note that more data (up to 50%) doesn't improve generalization further because of the overall unpredictability of the task, at least in certain conditions.

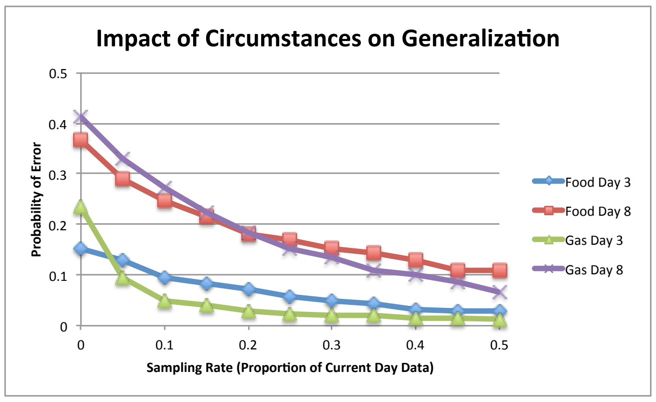


Figure 6: Effect of Circumstances on Spatial Model.

[3] We could refer to it as the integrated model because it also includes the temporal aspect through the base-level component, but we found the spatial/temporal distinction to be more descriptive.

Finally, to examine the impact of conditions on generalization, Figure 6 focuses on performance on Day 3 (2 days after the storm) and Day 8 (the day after a secondary storm hit) for food and gas outlets (pharmacy outlets omitted but results similar to gas). Because of the difficulty of predicting availability immediately after a disruptive event, the error rate is consistently and significantly higher on Day 8 than Day 3. However, as for the average across all days, performance significantly improves with sampling rate, becoming almost error-free on Day 3, which relies only on a single day of useful complete data (Day 2) and the specified proportion of the current day's data.

Discussion

Gu et al. (2014) applied a variety of algorithmic approaches to the prediction problem using this data set. They similarly differentiated their approaches between spatial and temporal algorithms. Their algorithms can be seen as specialized version of the cognitive mechanisms used here, e.g., the LastKnownState algorithm is simply the recency component of base-level activation without frequency, while the BestProxy algorithm is effectively partial matching without blending (or stochasticity). Recognizing the need to reflect both temporal and spatial data, they develop an algorithm that combines the best of the two approaches, in a way similar to, but more limited than, how those factors are combined in chunk activation.

The compelling argument for cognitive models, however, is not that they outperform a given machine learning algorithm. Rather, it is that they provide a way to augment human cognition in a way that is fundamentally compatible with it, for example by selecting a limited set of data to provide to the human decision maker that would result in the best human performance. In ongoing work, we are exploring mechanisms to introspect into the mechanisms of our cognitive model to drive data selection that would maximize its performance. We plan to then verify the model's predictions by collecting data in situations that combine model data selection and human decision making.

Acknowledgments

This research was supported by the Network Science Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory or the U.S. Government.

References

AI Magazine (2016). Beyond the Turing Test. AAAI Press. April, 2016.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y . (2004). An integrated theory of the mind. *Psychological Review* 111, (4). 1036-1060.

Cowan, N. (2001). "The magical number 4 in short-term memory: A reconsideration of mental storage capacity". *Behavioral and Brain Sciences* **24** (1): 87–114; discussion 114–85.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S., Hau, R., Hertwig, R., Stewart, T., West, R., Lebiere, C. (2010). A choice prediction competition, for choices from experience and from description. *Journal of Behavioral Decision Making 23(1)*: 15-47.

Gu, S., Pan, C., Liu, H., Li, S., Hu, S., Su, L., Wang, S., Wang, D., Amin, T., Govindan, R., Aggarwal, C., Ganti, R., Srivatsa, M., Bar-Noy, A., Terlecky, P., & Abdelzaher, T. (2014). Exploitation Data Extrapolation in Social Sensing for Disaster Response. The *10th IEEE International Conference on Distributed Computing in Sensor Systems* (DCOSS 2014), Marina Del Rey, CA.

Kurzweil, R. (2006). *The Singularity is Near.* New york, NY: Viking Press.

Lebiere, C. (1999). The dynamics of cognitive arithmetic. *Kognitionswissenschaft* [Journal of the German Cognitive Science Society] Special issue on cognitive modelling and cognitive architectures, D. Wallach & H. A. Simon (eds.)., 8 (1), 5-19.

Lebiere, C., Gray, R., Salvucci, D. & West R. (2003) Choice and Learning under Uncertainty: A Case Study in Baseball Batting. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*. pg 704-709. Mahwah, NJ: Erlbaum.

Lebiere, C., Pirolli, P., Thomson, R., Paik, J., Rutledge-Taylor, M., Staszewski, J., & Anderson, J. R. (2013). A Functional Model of Sensemaking in a Neurocognitive Architecture. *Computational Intelligence Neuroscience.*

Marr, D.; Poggio, T. (1976). "From Understanding Computation to Understanding Neural Circuitry". Artificial Intelligence Laboratory. A.I. Memo. Massachusetts Institute of Technology. AIM-357.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63 (2): 81–97.

Oaksford, M. & Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning.* Oxford University Press, Oxford, UK.

Stocco, A., Lebiere, C., & Samsonovich, A. V. (2010). The B-I-C-A of biologically inspired cognitive architectures. *International Journal of Machine Consciousness,* 2(2), 171-192.

Wallach, D., & Lebiere, C. (2000). Learning of event sequences: An architectural approach. In Proceedings of *International Conference on Cognitive Modeling* 2000, pp. 271-279. NL: Universal Press.

West, R. L., & Lebiere, C. (2001). Simple games as dynamic, coupled systems: Randomness and other emergent properties. *Journal of Cognitive Systems Research*, 1(4), 221-239.