Modeling the confidence of predictions: A Time Based Approach

Uwe Drewitz (uwe.drewitz@tu-berlin.de)
Department of Cognitive Psychology and Cognitive Ergonomics, Berlin University of Technology,
Franklinstr. 5-7, 10587 Berlin, Germany

Manfred Thüring (manfred.thuering@tu-berlin.de)
Department of Cognitive Psychology and Cognitive Ergonomics, Berlin University of Technology,
Franklinstr. 5-7, 10587 Berlin, Germany

Abstract

Everyday life demands explanations and predictions from everybody all the time. Using experience based knowledge, the human mind is well suited to draw the required causal inferences. However, due to failures in the past, such inferences are usually drawn under uncertainty and come along with different degrees of confidence. We present an ACT-R model describing the cognitive processes of induction and deduction for a prediction task in a simple, simulated technical environment. While ACT-R provides excellent mechanisms to capture causal learning and causal inferences, no process has been defined yet to account for the trust humans put in their predictions. Based on the availability heuristic by Tversky and Kahneman (1973), we propose an approach for modeling different levels of trust by using a temporal module from Taatgen, van Rijn and Anderson (2007), thus relating availability to retrieval time and confidence judgments. The forecasts of our model are compared with the results of an empirical study and nicely fit the experimental data.

Keywords: causal models; uncertainty; inductive learning; availability heuristic, temporal module; time estimation.

Introduction

The explanation of a current state of the world by events in the past and the prediction of future events from a present situation are fundamental qualities of human cognition. We follow the assumption proposed by many others that such reasoning processes are based on causal models (e.g., Waldmann, 1996) and proceeded under uncertainty (e.g. Einhorn & Hogarth, 1982). Two factors determine how much trust we put in an explanation or a prediction.

The first factor is the perceived amount of missing information in a given situation. This case applies when a causal model demands more data than currently available. Experiments by Thüring and Jungermann (1992) as well as Jungermann and Thüring (1993) demonstrated that such situations appear as ambiguous and lead to a reduction of confidence people have in their causal inferences.

The second factor is not an attribute of the situation, but of the causal model itself. Causal models – as any other kind of mental model – may be incomplete or even incorrect (Norman, 1983), hence leading to faulty conclusions.

Obviously, deficient models are not trustworthy. Confidence requires success, i.e., "…it's the *model's ability to make accurate predictions that is the ultimate measure of the model's value"* (Chown 2006, p. 69). This value can be characterized as the reliability of the model. To summarize, the ambiguity of the situation at hand and the reliability of the causal model currently employed determine the strength of confidence we have in the conclusions we draw. If we want to predict this confidence, we require a formal basis for modeling the influence of both factors. In the former studies by Thüring and Jungermann, rule-based systems served as such a basis and were used to describe the structure of a causal model. This approach was well suited to characterize ambiguous situations by the degree of matching between data and the conditional parts of the rules and to predict the content and confidence of causal inferences drawn from them. The reliability of a model, on the other hand, proved as more complicated to handle. Especially when we tried to describe how rules are formed in the course of inductive learning and which psychological mechanisms influence the confidence of causal judgments based on such rules "under construction", it became apparent that a comprehensive cognitive framework is needed to cope with the complexity of the matter.

The cognitive architecture ACT-R (Anderson, Bothell, Byrne, Douglass, Lebiere & Qin, 2004) provides such framework. We will use it to demonstrate how simple rule-based causal models can be built from induction and how predictions can be derived from such models. Special emphasis will be placed on the issue of how the success (respectively failure) of predictions in the course of learning influence the reliability of the rules and the confidence people place in their inferences.

Modeling Objectives

To model induction, predictions and confidence, three basic objectives must be achieved.

(i) To ensure inductive learning, not only the current situation must be represented in the ACT-R model, but preceding situations must be accounted for as well. In addition, the success or failure in coping with these situations must be captured. (ii) The ACT-R model must be able to make predictions. A prediction can be characterized as a statement about a future state of the world in terms of specific propositions. Since predictions are made under uncertainty, the ACT-R model must be able to combine a propositional content with a degree of confidence. To achieve this, reliability as well as ambiguity must be considered by the ACT-R model (although the latter is not emphasized here). (iii) In case of incorrect predictions, the ACT-R model must provide mechanisms to modify the causal knowledge structure if new evidence is available. To put the objectives into practice and to implement an ACT-R model with the ability to generate predictions with different

degrees of confidence, we have to refer to experimental data.

The Experiment

The empirical basis of our approach are data obtained in an experiment by Thüring, Drewitz and Urbas (2006) that tested the following assumption: When a causal model is induced from observations, inferences *deduced* from that model are usually probabilistic and their uncertainty is influenced by the observer's former experience with the model. The results of this experiment were extensively discussed in Thüring et al. (2006) to clarify the interplay of induction, deduction and confidence judgments.

In the experimental task, the participants had to acquire the causal model of a technical system, i.e., the cooling system of a power plant. The system could run properly (state OK) or not (state MALFUNCTION) and consisted of four pumping devices (subsystems A, B, C and D). Information about the subsystems was displayed on four dials (Fig. 1) which could be turned on (A, D) or off (B, C). Each dial represented the state of a subsystem that was either 'up' (A), 'down' (D), or 'unknown' because its dial was switched off (B, C). While each of the factors A, B and C was causally relevant at some point of the experiment, factor D was a random variable serving as a distractor, which was introduced to obtain a sufficient level of task complexity. In each trial, participants were shown a combination of dials as in the left part of figure 1. Based on this information, they first predicted the state of the overall system by pressing one of two buttons 'OK' or 'MALFUNCTION', and then rated their confidence by adjusting a slider. After submitting their confidence rating, a status message informed them about the correct system state as shown in the right part of figure 1.

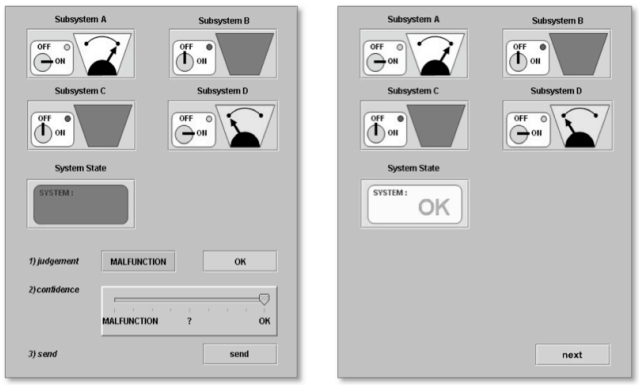

Figure 1: Screen layout of the experiment.

Using the feedback they received in each trial, participants could gradually develop a causal model representing the relation between the state of the subsystems (A, B, C, D) and the state of the entire system (OK or MALFUNCTION). In the first phase of our study, a simple model was induced in which just the proper functioning of one subsystem (e.g., A) was required for the faultless running of the cooling. Our participants learned this model fairly quickly from the data. In figure 8, the curve labeled "human" shows their mean confidence ratings (transformed into percentage values). Data points in the upper half of the figure represent ratings for the prediction "OK", those in the lower part for the prediction "MALFUNCTION". Note that the ratings start well above zero, because three trials in which A was coupled to OK were used in advance to acquaint the participants with the experimental setting. Starting from there participants soon reached a high and stable level of confidence (i.e., mean values between 70% and 80% with some exceptions due to the random condition D). At the end of this learning phase, information was provided which reduced the reliability of the model, i.e., in the trials 26-31 the feedback was contrary to the initial system behavior. Consequently, our participants' confidence in their predictions dramatically decreased and some of them even predicted a state contradictory to the rule they had learned before.

In the second phase of the experiment, information was provided that allowed for expanding the simple *'mono causal'* model into a more extensive one. This was either an '*or*-model' capturing multiple alternative causes each of them being sufficient for the effect, or an '*and*-model' representing a conjunction of several causal conditions each of them being necessary for the effect. When the new model was reinforced over several trials, confidence ratings raised to a level similar to the one of the mono causal model at the beginning (see fig. 8 and 9). When the reliability of these models was reduced (trials 31-35 and 45-49), the same effects occurred as in phase one, i.e., confidence ratings dropped again.

According to our first objective, the ACT-R model must be able to capture the cognitive processes of knowledge acquisition in this experiment, which are distinguished by the fact that people revise and expand their causal model when new facts become available.

Knowledge Acquisition

We propose three mechanisms of knowledge acquisition complementing each other, with each of them being necessary to form and diversify a causal model.

Inductive Learning

The first mechanism can be characterized as inductive learning. Within their natural environments, people make observations and store them in memory. Observing the same constellation of events repeatedly strengthens their associative relation in the memory trace. Thus, rudimentary causal models are constituted that guide further observations. In our experiment, these models could be described in terms of simple rules such as "if A is up then the system is OK" or "if A is down then a MALFUNCTION occurs".

Deductive Reasoning

Inductive learning is closely related to deductive reasoning. When a rule has been formed via induction, its reliability is

tested via deduction, thus creating a circle in which these two mechanisms take turns in forming a causal model. In each deduction, available data are matched with the rules and a conclusion is drawn. Those rules, which have been reliable in the past, are chosen over less reliable ones. In the first phase of our experiment, the rule "if A is up then the system is OK" produced a correct prediction whenever A was up, while the rule "if D is up then the system is OK" did not, because the relation between D and the system state was random.

Though reliable rules should be chosen most frequently, less reliable ones can get a chance when their conditions are matched by the current data. When this happens, the confidence that is placed in the prediction should be less compared to the confidence in a prediction derived from a reliable rule. For example, the confidence in predicting a well functioning system when "D is up" should be lower compared to a situation when "A is up".

To summarize, reliability serves two purposes. It determines which rules are chosen over others and it tunes the confidence people place in their predictions. Both these functions must be implemented in ACT-R to explain the data of our experiment and to achieve the second objective stated above, i.e., the derivation of the propositional content of a prediction in combination with a specific degree of confidence based on the experienced reliability of the model.

Rule Revision

While the reduction of confidence placed in a prediction is one consequence of the failure of a rule, the revision of the rule itself is another one. Changing the content or the structure of a rule is the third mechanism required to describe the forming of a causal model. Revisions only make sense in the light of new evidence, i.e., when the failure of a rule coincides with the observation of new conditions that must be satisfied in addition to (or instead of) the conditions that have been accounted for so far. In this case, the rule in question is altered. In our experiment, this happened in the second phase where simple mono causal models where expanded to an "or-model" or an "and-model". To attain our third objective, such changes must be accounted for when causal models are developed in ACT-R.

Overview of the ACT-R Model

The three mechanisms were implemented in the framework provided by ACT-R 6.0. Figure 2 displays the cyclic concept we used to establish the cognitive flow of control for performing the successive trials in our experiment. The nodes represent different control states, whereas the directed links indicate possible transitions between them.

At the START of each experimental trial, the current situation is stored in an ACT-R buffer. This situation consists of the states ("up" or "down") of the four components (A to D) of the cooling system. The task is to predict if these states will entail a proper functioning or a malfunction of the system. The next step is to SEARCH for instances in declarative memory matching the situation at

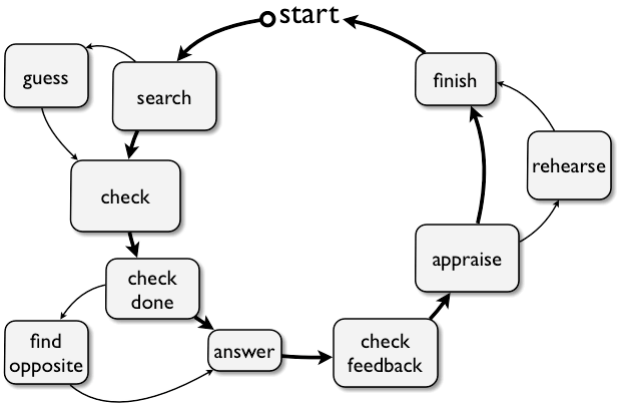

Figure 2: Cognitive flow of control of the ACT-R model.

hand. In our model, each search in memory relies on the spreading activation mechanism and is affected by noise resulting from the according parameter in ACT-R. Those instances, that have been frequently used in former cycles, have a higher activation and hence a higher probability to be found. Two outcomes are possible at this stage. (i) If a match is made, the according instance is retrieved. Now, a first propositional content for the required prediction has been found, since the instance contains the effect (OK or MALFUNCTION) that this specific constellation of A to D has produced in the past. To account for previous experiences with the prediction, its content is linked to an appraisal value. The appraisal is "good" when former predictions were correct, but "bad" when mistakes were made in the past. (ii) If no match is made, the model switches to "GUESS". This is the case either when the current situation is new, or if the activation of no instance in declarative memory is high enough for a successful retrieval. Guessing means that one of the two outcomes "OK" or "MALFUNCTION" is chosen at random from declarative memory. Therefore, in case (i) as well as in case (ii), the result at this stage is a first propositional content for the required prediction enhanced by an appraisal value.

In the next step, the content is CHECKED against different experiences made in the past. Three alternatives are possible: (i) If an instance with "good appraisal" was found during SEARCH, the check looks for an instance with the same effect but a "bad appraisal". (ii) If an instance with "bad appraisal" was found during SEARCH, the check looks for an instance with the same effect but a "good appraisal". (iii) If the result of GUESSING was "OK", then the alternative effect "MALFUNCTION" is produced during memory search, and vice versa for the result of "MALFUNCTION". The idea underlying this stage is twofold. First, it mimics reasoning under uncertainty where inferences are compared to other possibilities. Second, the cognitive processes involved here produce different retrieval times that are used to model different degrees of confidence. How this is achieved will be described later.

When the CHECK has been accomplished, the model switches to CHECK DONE. Now, if the appraisal of the instance retrieved during SEARCH is "bad", the preliminary

propositional content is not reliable. In this case, the process FIND OPPOSITE generates the opposite effect as alternative prediction and uses it as ANSWER. Otherwise, no alternative prediction is required and the result of the former SEARCH is delivered as ANSWER. In our experiment, this is the point where subjects make their prediction and then receive a feedback.

In the ACT-R model, the FEEDBACK is CHECKED by comparing it with the prediction. If the prediction is correct, APPRAISE generates the appraisal value "good" and links it to the according instance. In case of a wrong prediction, however, the appraisal turns to "bad" and hence the instance represents an incorrect prediction.

If a successful prediction is made based on GUESSING or on TAKING THE OPPOSITE, this new information is REHEARSED to strengthen the activation of this valuable new insight. For the same reason, REHEARSAL occurs when a formerly reliable instance produces a wrong prediction.

When the state FINISH is reached, all buffers are cleared and the results are transferred to declarative knowledge. A result consists of a new instance, whenever an unprecedented constellation was encountered in that cycle and used for a prediction. In this way, declarative knowledge is extended and revised.

So far, we have described a circular process of knowledge acquisition consisting of inductive learning, deduction and rule revision. Figure 3 shows the predictions made by the ACT-R model (over 21 runs) compared to the predictions made by the participants in the experiment by Thüring et al. (2006). As indicated in the chart, there is a very good fit between both types of predictions.

To fulfill our second objective, these results must be related to the generation of confidence judgments. Reliable causal rules are represented by instances with a positive appraisal. Among these instances, those with a high activation constitute a person's actual causal model. The amount of activation not only determines which rules are used for prediction, but should also influence the confidence people have in their predictions.

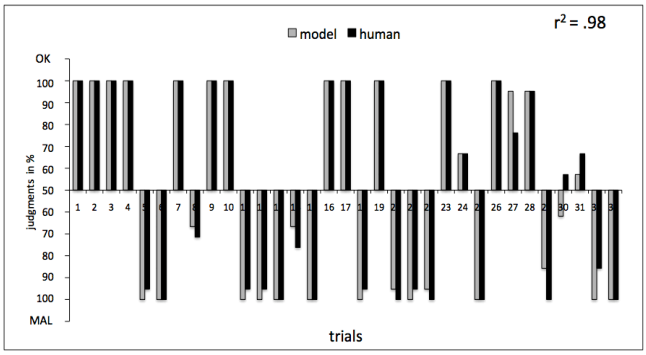

Figure 3: Mean propositional judgments (n=21).

However, since activation is a subsymbolic parameter, it cannot be directly used to produce confidence judgments. To solve this problem, we adopt a heuristic proposed by Tversky and Kahneman (1973) to our ACT-R model.

The Availability Heuristic: Degree of Confidence and Retrieval Time

When people have to evaluate the frequency or likelihood of an event, they often use heuristics to do so. In case of applying the availability heuristic, the subjective probability of an event depends on how fast the representation of a former occurrence of the event can be retrieved from memory, i.e., the faster the retrieval of the event, the higher its estimated probability. Tversky and Kahnman (1973) assumed that the ease of retrieval is equivalent to the perceived time of retrieval. This offers an interesting solution for the problem of modeling the confidence of predictions. The retrieval of an instance raises its overall activation, which in turn lowers its retrieval time and hence should increase the confidence in its propositional content. Within ACT-R, the perception of time can be captured by a temporal module that was developed by Taatgen, van Rijn and Anderson (2007), especially for estimating short times.

Estimating Time with the Temporal Module

The temporal module consists of a pacemaker and its relations to a temporal buffer (see fig. 4). Three different parameters can be set to influence time estimation within this framework (Taatgen et al., 2007). One of them is the *time-master-start-increment*. This parameter has to be set at a low level to make the module sensitive enough for estimating short durations, such as retrieval times.

When time measuring begins, a start signal is created which causes the pacemaker to generate time pulses, so-called ticks. These ticks are collected in the accumulator of the temporal buffer. When a time estimation is needed, the number of ticks that have been accumulated between the temporal request and the retrieval represents the elapsed time.

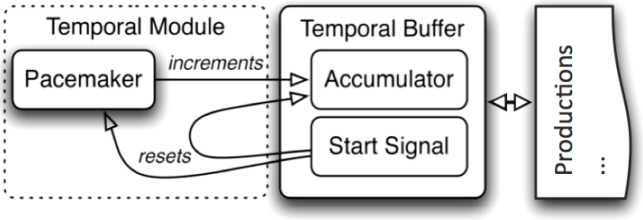

Figure 4: The temporal module (taken and adapted from Taatgen et. al., 2007).

In our approach, time estimation is always related to the retrieval of a specific memory element, such as an instance. Therefore, any temporal request is combined with the request for a memory element, and the analog holds for the retrieval. The ACT-R syntax implementing the combined request and retrieval is shown in figure 5.

The result of a temporal retrieval is a symbolic value characterizing the perceived time for finding the memory element. This value can be processed further to generate different degrees of confidence.

```
1. Combined Request
(P start-request
...
+retrieval>
      isa memory-element
+temporal>
      isa time
...)
```

```
2. Combined Retrieval
(P harvest-retrieval
...
=retrieval>
      isa memory-element
=temporal>
      isa time
      ticks =ticks
...)
```

Figure 5: Combined declarative and temporal request.

From Time to Confidence

We propose two different methods to transform perceived retrieval times into confidence judgments. Both are mathematical functions, which (at least for the time being) are not implemented within ACT-R itself.

Transforming retrieval time. The first method can be characterized as a direct implementation of the availability heuristic. It is expressed by the formula in figure 6. Two properties of this function are immediately salient: (i) Short retrieval times lead to high confidence values while long retrieval times cause low confidence judgments. (ii) Since the function is logarithmic, the decrease of confidence decelerates with the number of ticks increasing. This accounts for the observation that differences between longer retrieval times result in rather small differences for related confidence ratings and vice versa.

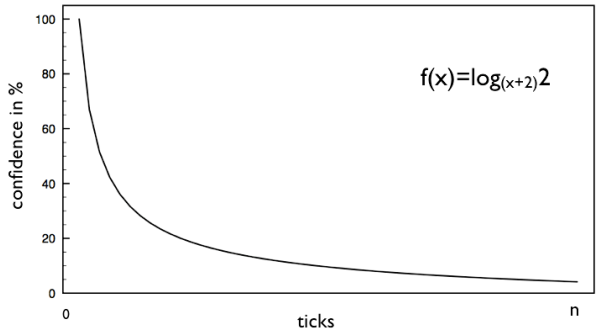

Figure 6: Transformation function ($f(x)=\log_{(x+2)}2$) for the transformation of *retrieval time* (schematically).

Figure 8 displays the confidence judgments for predicting the system states "OK" and "MALFUNCTION" that are generated by our model when this function is used. Although the match between the model and human data is good, a more sophisticated approach can be taken to model the confidence of predictions.

Transforming retrieval time differences. The idea underlying our second method is to check the retrieval time for an original prediction against the retrieval time for an alternative prediction. The alternative is an instance of the same content, but with an appraisal indicating that (at least once) the instance has failed to be successful. Due to its success in the past, the original predication is highly activated and can be retrieved fast. If the same holds for the alternative, the difference between the retrieval times of both predictions is small and the confidence in the original should be low. On the other hand, if the alternative prediction has been less successful than the original, its lower activation entails a longer retrieval time. In this case, the difference between the retrieval times of both predictions is large and the confidence in the original prediction should remain high. These relations between retrieval time and degree of confidence are captured by our second function. It accounts for the fact that we may find conflicting information of different value when we search our memory to make a prediction.

The difference of both retrieval times is calculated (as an absolute integer) and taken as input for the transformation process. Figure 7 shows the formula and form of the function used for this transformation.

Figure 9 presents the generated data by the model using the method of transforming time differences into confidence ratings.

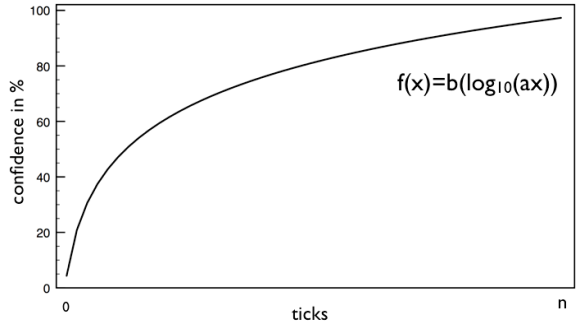

Figure 7: Transformation function ($f(x)=b*\log_{10}a*x$) for the transformation of *time differences* (schematically).

Discussion

The ACT-R model and the two functions described above were developed to account for the data of the first experimental block where a *'mono causal model'* had to be learned. A comparison of the charts in figure 8 and 9 indicates that both functions are well suited to model confidence ratings based on time measures.

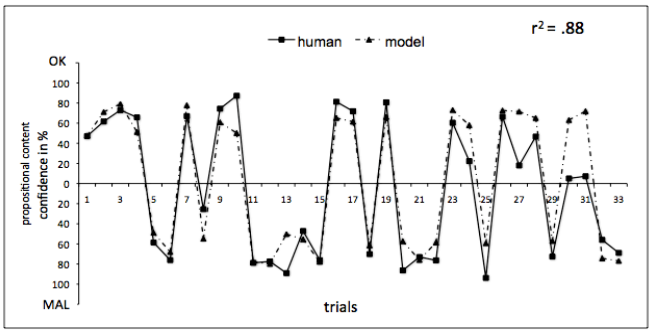

Figure 8: Combined ratings for the *'mono causal'* block calculated with method I: transformation of retrieval time (n=21, RMSSD=4.3).

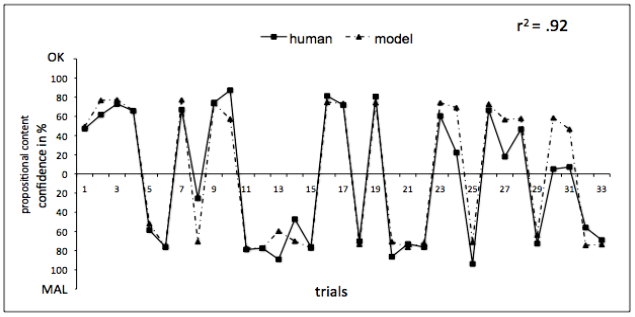

Figure 9: Combined ratings for the *'mono causal'* block calculated with method II: transformation of time differences (n=21, RMSSD=3.1).

Nevertheless, there is an advantage for the second function. The trend measure (r^2) and the goodness-of-fit measure (RMSSD) show a better fit with the empirical data for that method. Therefore, the second function was chosen to predict the confidence ratings in the second experimental block, where the *'and-model'* as well as the *'or-model'* were induced. Again, the model proved to be well applicable, matching the empirical data with a high fit (see fig. 10 and 11).

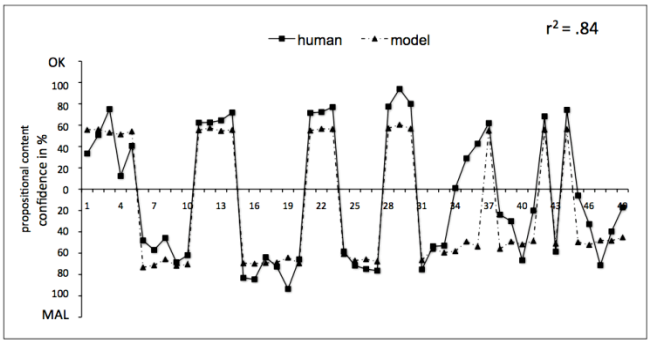

Figure 10: Combined ratings of propositional content and related confidence for the *'and'* block (n=21, RMSSD=4.3).

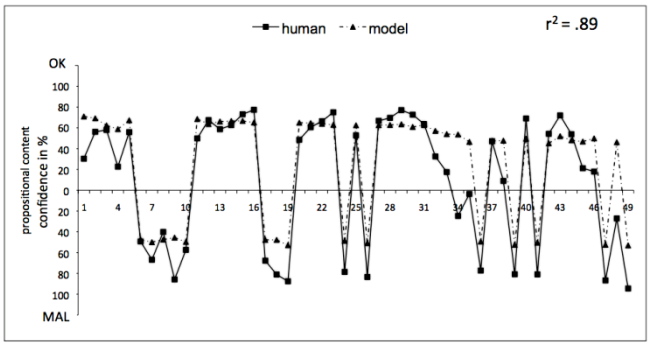

Figure 11: Combined ratings of propositional content and related confidence for the *'or'* block (n=21, RMSSD=3.8).

To summarize, we have proposed an ACT-R model, which combines inductive learning, deductive reasoning and mechanisms for revising knowledge structures to describe the acquisition of causal models. Predictions derived from a causal model are made under uncertainty, i.e., the propositional content of an inference is combined with a particular confidence. In order to describe different degrees of confidence, the availability heuristic proposed by Tversky and Kahneman (1973) was adopted to our ACT-R model. This was accomplished by using estimated retrieval times of memory elements to operationalize availability. The operationalization was achieved by two mathematical functions, which transform retrieval times into confidence judgments. The data generated by our ACT-R model in combination with these functions where compared to data generated by humans in an experiment reported by Thüring et al. (2006). As a result, the second function proved as slightly superior to the first one.

Future research will adress the problem of how this function can be implemented directly within the ACT-R framework. Moreover, our approach must be tested in further experiments adressing different situations of inductive learning as well as different domains of reasoning.

References

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.

Chown, E. (2004). Cognitive Modeling. In A. Tucker (Ed.), *Computer Science Handbook*. Chapman & Hall.

Einhorn, H. J. & Hogarth, R. M. (1982). Prediction, diagnosis, and causal thinking in forecasting. *Journal of Forecasting*, 1, 23-36.

Jungermann, H. & Thüring, M. (1993). Causal knowledge and the expression of uncertainty. In G. Strube & K. F. Wender (Eds.), *The cognitive psychology of knowledge*. Amsterdam: Elsevier.

Norman, D. A., (1983). Some Observations on Mental Models. In D. Gentner & A.L. Stevens (Eds), *Mental Models*. NJ: Hillsdale.

Taatgen, N. A., Rijn, H. v., & Anderson, J. R. (2007). An Integrated Theory of Prospective Time Interval Estimation: The Role of Cognition, Attention and Learning. *Psychological Review*, 114(3), 577-598.

Thüring, M., Drewitz, U., & Urbas, L. (2006). Inductive Learning, Uncertainty and the Acquisition of Causal Models. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Cognitive Science Society*. NJ: LEA.

Thüring, M. & Jungermann, H. (1992). Who will catch the Nagami Fever? Causal inferences and probability judgment in mental models of diseases. In D. A. Evans & V. L. Patel (Eds.), *Advanced models of cognition for medical training and practice* (pp. 307-325). Berlin: Springer.

Tversky, A. and D. Kahneman, 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* (5), 207-32.

Waldmann, M., (1996). Knowledge-based causal induction. In D. R. Shanks, D. L. Medin & K. J. Holyoak (Eds), *The Psychology of Learning and Motivation (Vol. 34): Causal Learning*. San Diego, CA: Academic Press.