Memory and Contextual Change in Causal Learning

Uwe Drewitz (uwe.drewitz@tu-berlin.de)
Stefan Brandenburg (stefan.brandenburg@tu-berlin.de)
Department of Cognitive Psychology and Cognitive Ergonomics, Franklinstr. 28/29
10585 Berlin, Germany

Abstract

Declarative memory is a central resource for reasoning processes. In line with the ACT-R theory, we assume that declarative memory is the basis for causal learning. Based on this assumption we conducted an experiment, showing that subjects' confidence in causal predictions decreased if their causal knowledge is discredited. Moreover, confidence decreased not only for the causal knowledge that was discredited, but also for knowledge that was not at all manipulated. Additional to the experimental results, we present an ACT-R model that perfectly fits the data and provides an explanation for the empirical findings. Contextual change turns out to sufficiently explain the empirical data and the principle of our ACT-R model.

Keywords: Contextual change, activation, causal knowledge, inference.

Introduction

The central role of memory has been investigated in a widespread range of tasks. There is much evidence showing that especially declarative memory accounts for human performance usually seen as smart or intelligent behavior (e.g. Anderson, 2007). We assume that causal learning and causal reasoning is largely based on declarative memory as well. This assumption is in line with recent research on reasoning (Mehlhorn, Taatgen, Lebiere & Krems, 2011) and the application of heuristics (e.g. Schooler & Hertwig, 2005). This research was and still is based on the ACT-R theory (Anderson,Bothell, Byrne, Douglass, Lebiere & Qin, 2004). In ACT-R human declarative memory is responsible for the storage of factual information. This information is stored in chunks. These chunks become available for retrieval, based on their activation. The higher the activation, the higher is the probability of retrieval and the faster is the retrieval of a chunk. This is the central functional principle of declarative memory in ACT-R. The activation of a chunk reflects both, the history of its usage as well as its relevance for the current context. Both aspects of activation are relevant for the explanation of human performance.

Decision-making under uncertainty is an example where human performance relies on declarative memory (Tversky & Kahnemann, 1974; Hertwig, Herzog, Schooler & Reimer, 2008, Gigerenzer & Gaismeier, 2011). Human behavior in such situations can be explained by retrieving instances of memory. However, peoples' performance cannot be explained by the mere retrieval. Instead in literature principles are proposed, which are related to the retrieval. First, in the availability heuristic (Tversky & Kahnemann, 1973) subjects evaluate how available or how accessible (Kahnemann, 2003) a memory chunk is. Second, Schooler and Hertwig (2005) propose that people evaluate the difference in retrieval times for alternatives. This research assumes those peoples' confidences ratings in decision-making under uncertainty dates back on these by-products of the retrieval process. Also for causal learning, Drewitz and Thüring (2009) showed that peoples empirical data can be explained based on the interpretation of retrieval times. As ACT-R frames retrieval times as dependent on activation it can be concluded that confidence of ratings are directly related to the activation of memory elements. But this claim holds only for performances and experiences solely based on memory retrieval. To conclude, from the ACT-R point of view these performance and confidence ratings are explained by activation. The model results presented in this paper give evidence to this position.

Sufficiency and Necessity in Causality

It has been proposed, that human causal learning relies on cues to causality (Einhorn & Hogard, 1986). One of these cues is the co-variation between events, which people can obtain from contingency data. Theoretical approaches that emphasize the role of covariation assume that in causal learning and reasoning persons rely on frequencies of (co-)occurrence and (co-)absence of event. Figure 1 shows how this contingency information can be depicted for two events. The four cells represent the four possible pairings of two events (*C* and *E*). With respect to these two events, every observation can be assigned to one pairing and as such, to one cell of the contingency table. Moreover, every observation gives either positive or negative evidence to one of two aspects of causality: *sufficiency* and *necessity*. People are willing to attribute a causal relation between two events if both aspects are met. John Stuart Mill (1869) first made this claim.

According to him, people don't acquire causal knowledge from the repeated observation that one event follows the other. Instead they take into consideration what happens if a putative cause does not occur. From this perspective, causes can be characterized in terms of *sufficiency* and *necessity* and both of these aspects have to be satisfied. And they are satisfied to the full extent, if a number of observations fall into cell a and d as well, but not in cells c or b. In other words, every observation that belongs to the event pairing of cell a gives positive evidence to the *sufficiency* of the putative cause C for E, the effect of interest. Just as all observations that belong to the pairing of cell d give evidence to the *necessity* of C for E.

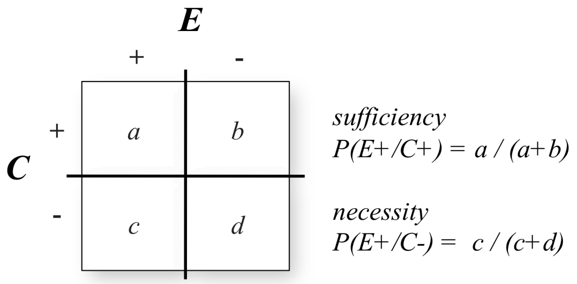

Figure 1. 2x2 contingency table ('+' indicates presence, '-' indicates absence).

Moreover, *sufficiency* and *necessity* are statistically independent of each other. Whereas the *sufficiency* of C for E depends on the frequencies in cells a and b, the *necessity* is determined by the frequencies in cells c and d (see Fig.1). Two different conditional probabilities capture these facts (see Fig. 1): the probability of the presence of *E* given the presence of *C*, P(E+/C+), and the probability of the presence of *E* given the absence of *C*, P(E+/C-). Positive evidence for one aspect (observations that fall either into cell a or d) can be understood as strengthening an aspect. Comparably, negative evidence (observations that fall either into cell b or c) weakens one of both aspects. Theories that emphasize the role of co-variation as cue to causality (for review see Perales and Shanks, 2007), describe how people integrate their knowledge about both aspects. That seems to be important especially when participants in a causal learning / reasoning task are requested to rate the strength of a causal relation. Of course, people do integrative judgments like that also in real-world tasks. But very often they for example make predictions based on data. In turn, as soon as people can rely on e.g. C+, their prediction should be related only to sufficiency of *C* for *E* (cells a and b). In such a case, there is no need to integrate the information about the opposite i.e. C-, which is captured by the frequencies in cells c and d. This is also true the other way around. To sum up, for the predictions based on given data, there is no need to integrate information that would hold for the absence of that data. Consequently, given the independence of both aspects, neither positive nor negative evidence related to one of the aspects should affect inferences related to the complementary aspect. Standard theories (see Perales and Shanks, 2007) do not propose such an effect. In contrast, we claim that such an effect is there. The underlying assumption is, that people do not render *sufficiency* and *necessity* as independent as they are from a mathematical point of view. Moreover, we assume, that they treat them as belonging together. And in fact, as complementary parts they belong together in terms of the concept of causality. With respect to observations people make in the world, both parts are summing up to a bigger whole – our knowledge of causal relations. But if people treat them as parts, which shape together as a whole, it can be assumed that if one part fails, people do not longer trust in the other part. In turn our hypothesis states that the impact of negative evidence for one aspect of causality is twofold. First of all it weakens the aspect that was *discredited* by negative evidence. As a result peoples confidence for predictions related to that aspect would drop down. Thüring, Drewitz & Urbas (2006) showed this effect. Second, the complementary aspect, i.e. the aspect that is not discredited will be *devaluated.* That means peoples confidence for predictions to that aspect will decrease as well. This would be in contrast to the fact that sufficiency and necessity are independent of each other. We tested this hypothesis in our Experiment.

Basic Causal Models

With respect to the concept of causality building upon sufficiency and necessity Thüring & Jungermann (1992) proposed that people's representation of causal knowledge could be described in terms of causal models. There are different basic causal models, which can be used to represent basic or if combined, more complex causal relations.

The building blocks of all these models are conditional rules. For every causal relation the aspect of sufficiency as well as the aspect of necessity is captured by one or more of these rules. In Table 1 the sets of rules for two basic causal models are shown. These are the models that are addressed in our experiment: the *model of unique causation* and the *model of compound causation*.

Table 1. Basic causal models.

model of unique causation	model of compound causation
R1: C+ → E+	R3: C+ and X+ → E+
R2: C- → E-	R2: C- → E-
	R4: X- → E-

Experiment

In our study, participants had to acquire causal knowledge about a simulated technical system based on inductive learning. Over the course of the experiment, positive as well as negative evidence was presented to investigate the consequences of discrediting and devaluation.

Method

Participants. Fifteen graduate and undergraduate students at the Berlin Institute of Technology were recruited for the experiment. All of them were paid for their participation.

Material. Figure 2 shows the schematic screen layout of the simulated system that was presented to the participants. It was introduced as an electrical system of a power plant. The system was built up from four subsystems that were responsible for two output systems. Information about the state of these subsystems was displayed on four dials (for top boxes in Fig.2). Each dial represented the state of one subsystem, which was either DOWN (C+) or UP (C-) or UNKNOWN because its dial was switched off. In the first of two blocks only one subsystem was causally relevant and its state served as cause (*C*) for the outcome (either E+ or E-) of the relevant output system (*E*). In the second block the same subsystems (C) together with another subsystem (X) was causally relevant. During the first block X was always set to UNKNOWN. The other two subsystems were irrelevant for the task. In both blocks they were used in some trials as distracters to give the system a more diversified appearance.

In the lower half of the screen, the displays for the output systems were shown. In some of the trials participants had to predict the outcome of only one of them and in the remaining trials they had to predict the outcome of both. If only the outcome of one system had to be predicted the display of the other output system wasn't shown. Whereas one output system (*E*) was relevant for the experiment the other was used to make the task more realistic. Below the display of each output system two buttons were shown for the prediction of the outcome (depicted as '+' and '-' in Fig. 2.) One button served the prediction of MALFUNCTION (E+) and the other one the prediction of OK (E-). Clicking on one of them was necessary to make the prediction. Finally, below these buttons a slider was presented (see Fig. 2) that could be adjusted to rate the confidence of the predictions. The lowest confidence (0%) was set in the middle of the slider, between the two maxima (100%), each related to one of the two possible outcomes.

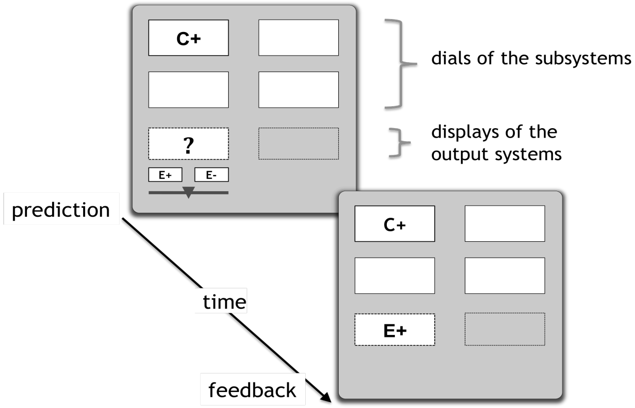

Figure 2. Screen layout (schematic) as used in the experiment for prediction and presentation of feedback.

Procedure. The participants' task was to predict the outcomes (E+ or E-) of the output system(s). To solve this task, they had to understand the underlying causal relation between the subsystems and the output systems.

In each trial, they were shown the layout of the device as presented in Figure 2. First, subjects had to check the operation of the subsystems. Based on this information, they were requested to predict the state of the output system(s) by clicking on the respective buttons (OK or MALFUNCTION). Finally, they rated their confidence for each prediction by adjusting the respective slider(s). After participants finished their prediction and confidence rating, they had to click on a 'send' button and subsequently received feedback that showed the actual outcome(s).

The experiment consisted of two blocks, each of them with a *learning phase* and a *test phase* as shown in Figure 3. The blocks differed in the complexity of the underlying causal relations and the number of trials in the learning phase. During the learning phase positive evidence for one type of causal relation was provided. In the learning phase of the first block participants received information that enabled them to acquire a model of unique causation with the two rules R1 and R2 (see Tab. 1). In the learning phase of the second block subjects received information, which supported the acquisition of the two new rules R3 and R4 (see Tab.1). Thus subjects could learn a new model, the model of compound causation (see Tab.1).

Additionally, we presented distracter trials with information about the irrelevant subsystems and trials were participants had to predict the outcome of the second output system that was irrelevant for the test of the hypothesis. Relevant for the test of the hypothesis were the pre-measure and the post measure in each block. For both blocks, the last trial of the learning phase served as *pre-measure* (see Fig. 3). In the respective trial in block one, people had to make a prediction based on C- and received as feedback E-. In the respective trial for block two we presented C+ and X+ and gave people after they made their prediction the feedback E+. These pre-measure trials served ass positive evidence too. That's why for both blocks in Figure 3, the number of one cell increases from the learning to the test phase.

Subsequently to the learning phase, the test phase started. In four of these trials, we presented negative evidence (see Fig. 3) for one aspect of causality. The negative evidence was given with respect to the causal relations that were supported before. You can see the number of presentations of negative evidence in the black boxes in Figure 3. In block one, we showed four times C+, E- and in block two we presented two times C-, E+ and two times X-, E+ (together four times negative evidence). Again, distracter trials and trials focusing on the irrelevant output system were presented. In the last trial of the test phase, the *post-measure* was recorded (see Fig.3). To accomplish this, the same data were given as in the pre-measure trial. For block one that was C- and for block two C+ and X+.

Independent and dependent variables. To investigate the *strengthening* of rules, the amount of *positive evidence* ranged from one to sixteen trials (see Fig. 3, positive evidence) for the respective rules (R1, R2, R3 and R4). To test the impact of discrediting, the amount of *negative evidence* ranged from one trial to four trials (see Fig.3, negative evidence) for the respective rules (R1, R2 & R4).

The factor *measurement* with the factor levels *pre* and *post* served the investigation of devaluation as described in the procedure (see Fig.3). Throughout the experiment, confidence ratings of inferences predicting the states of the relevant output system were used as dependent variable.

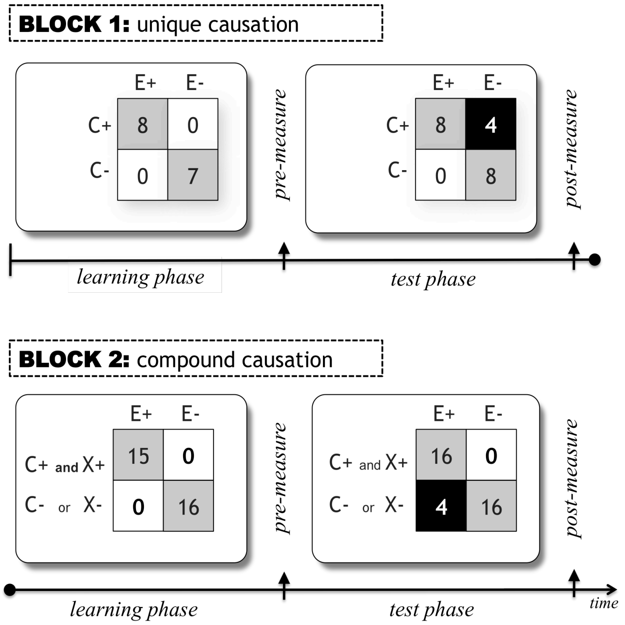

Figure 3. Experimental procedure (schematic). Contingency tables for both blocks display frequencies of positive evidence (grey) for the learning phase and negative evidence (black) for the discrediting phase. The last presentation of positive evidence served as pre-measure. The contingency tables on the right display summed frequencies for positive and negative evidence.

Results

For statistical analysis, we computed three ANOVAs with repeated measures, one for each effect. Additional to the significance of effects we report the effect size f after Cohen (1988). However, *strengthening* greatly affected subjects confidence ratings over the course of the learning phases, F(7,98)=11.47, p<0.01, f=0.91. Therefore, subjects' confidence in their prediction of the state of the output system strongly increased over time. Additionally, we obtained an effect for rule, F(3,42)=7.46, p<0.01, f=0.73. Hence, it was easier for subjects to acquire basic rules (R1 & R2) compared to the more complex rules (R3 & R4).

For *discrediting*, we found a significant large main effect of *negative evidence* over time, F(1,14)=4.21, p=0.05, f=0.54. Hence, subjects showed lower confidence in their ratings about rules that were discredited.

To investigate the effect of *devaluating* a rule, it seems necessary to highlight how we achieved the data for this computation. For all subjects R1 and R2 (in the first block) and R3 and R4 (in the second block) were strengthened. The last trial of R2 in the learning phase of block 1 and R3 in the learnigg phase of block 2 served as pre-measure for the devaluation. However, only R1 was discredited in the first block and R2/R4 were discredited in the second block. If subjects' confidence for the prediction of R2 in the first block (post-measure) and R3 in the second block (post-measure) was lower after discrediting the other rules, devaluation took place. The ANOVA revealed a medium main effect of *rule* (F(1,28)=5.42, p=0.03, f=0.43) and a large main effect of *measurement*, F(1,28)=19.40, p<0.01, f=0.83). Therefore subjects' confidence was lower for R2 compared to R3. Additionally, devaluation lead to significantly lower confidence for participants' confidence for both rules (R1 & R3) after R1, R2 and R4 were discredited.

Model Description

In order to examine, whether the empirical observed effect of devaluation could be explained based on declarative memory and the concept of activation, we set up a simple ACT-R model. Two central assumptions were made to specify the model. At fist, we assumed that the task could be processed solely based on instance retrieval.

The second assumption we made was in contrast to the theoretical considerations, which guided the formation of the hypothesis of the devaluation effect. To model the effect found in the empirical data, we assumed that negative evidence, which per definition contradicts observations made beforehand, would be considered as a contextual change (see Block and Reed, 1978). Accordingly, people are aware of changes in the context internally (cognitive context) as well as externally (external context). Assuming that observations, which people make, and certain knowledge that they acquire accordingly, is related to a certain context would result in a change of availability of that knowledge if the context changes.

Consequently, the behavior of the model can be explained based on contextual changes and instance retrieval, i.e. the standard ACT-R 6 activation equations:

$$A_i = B_i + \sum_i W_j S_{ji} + \sum_l PM_{li} \qquad \text{(activation equation)}$$

Accordingly, the Activation (A_i) of a chunk *i* is defined by its base-level activation (B_i), the amount of activation that spreads out from a source (representation of the stimuli and the current context) and the partial matching component. W_j reflects the attentional weigthing allocated to every element *j* on the source of activation. S_{ji} is the strength of the associative connections between these elements and the chunk *i*. *P* is the mismatch penalty and M_{li} is the similarity between the elements *l* specified for a request of retrieval from declarative memory and the respective elements of chunk *i.* The base-level B_i itself is defined as

$$B_i = \ln(\sum_{j=1}^{n} t_j^{-d}) \qquad \text{(base-level learning equation)}$$

where *n* is the number of presentations, *t* is the time that passed sine the *j*th presentation and *d* is the rate of decay. Last but not least the associative strength is defined as

$$S_{ji} = S - \ln(fan_j) \qquad \text{(associative strength equation)}$$

where S is the maximum associative strength and fan_j the number of chunks in memory, the chunk j as an element on the source of activation is associated with, plus one for association with itself.

Model Settings. Based on these equations the activation of that chunks was calculated, which would match the retrieval request in the trials of the relevant measures. Therefore we calculated the number of presentations (see base-level learning equation), assuming that there was one encoding of screen information as well as one declarative retrieval per trial. For simplicity reasons, the respective times (t_j) were calculated based on the assumptions that all trials were processed in the same time. To determine the associative strength between the current context and the different memory chunks (representing different trials) we counted the number of distinct stimuli used in the experiment. The respective number of associative connections was assigned to each context and so the fan for each context was set.

Since there is no default, one a single parameter, the mismatch penalty, was set to 0.5. Except this, all other parameters used (see equations), were set to their defaults prescribed by the ACT-R theory (cf. http://act-r.psy.cmu.edu/).

Subsequently, we transformed the resulting activation values (A_i) into values representing confidence ($confidence_i$) using the following equation:

$$confidence_i = \ln(A_i) \times SF \qquad \text{(transformation equation)}$$

The parameter SF is a scaling factor and was set to 100. The results produced by the model are shown together with the empirical data in Figure 4. Without adjusting any parameter the model produces an excellent fit to the empirical data (R^2=1). The qualitative match between the model results and the human data is perfect. Quantitatively there is a clear deviation. But this deviation is of same for all conditions.

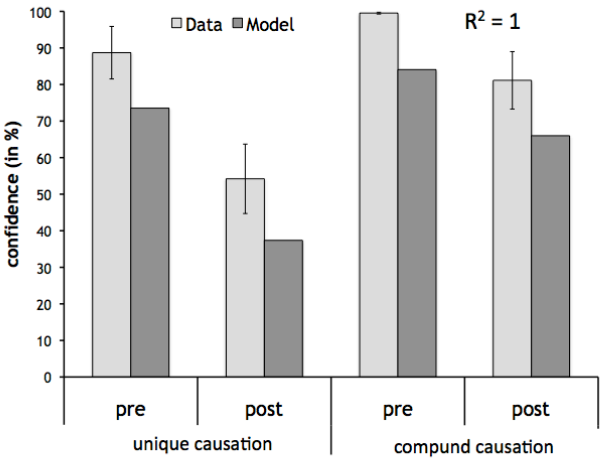

Figure 4. Empirical data and model results. Error bars represent standard error.

Discussion

The present paper investigated two effects that influence causal learning, memory and contextual change. In particular, we looked at persons' confidence ratings with respect to their predictions of certain outcomes after their causal knowledge was (a) strengthened and (b) discredited.

In line with previous research (Thüring et al., 2006), we found that the presentation of positive evidence leads to higher confidence in subjects' predictions of the state of the output system. Additionally, we replicated the effect of discrediting. Hence participants' confidence in their prediction of the state of the output system decreased, when the rules they had learned were discredited. In contrast to these well-established effects, we demonstrated the effect of devaluation. Therefore, participants' over-all confidence in their predictions of the system state decreased also when the rules they could use for prediction on given date were not discredited. This effect opposes the assumption that people treat sufficiency and necessity as independent as they are from a mathematical perspective. At least for the case of negative evidence the data support this position.

Excluding the striking effect of devaluation, memory effects (strengthening and discrediting) on confidence in causal learning were proposed and modeled in ACT-R before (e.g. Drewitz & Thüring, 2009). Extending this research, the present ACT-R model accounts for the effect of devaluation as well. For both causal relations (unique and compound causation), model data perfectly fitted to subjects' behavior in the experiment. It is important to note that this simple ACT-R model mimics empirical data about subjective confidences without any additional parameters. Hence the present data (and model) strongly support the theoretical claim we made at the beginning of this paper. There, we proposed that peoples' confidences of ratings under uncertainty are directly related to the activation of memory elements. The results presented in this research undermine this claim at least for performances and experiences, which are solely based on memory retrieval.

However, the presented ACT-R model does not only mimic the empirical data. Its working principle also provides an elegant theoretical way to explain the data. We used the concept of contextual change (Block, 1978) as basis for the ACT-R model. Contextual change means that with respect to their model or rule like causal knowledge people consider certain contexts. In our experiment (and model), subjects learned causal relations in each block. After a couple of trials, their rule-based knowledge about the functioning of the technical operating system was almost perfect. This phase of strengthening contained only positive evidence. Therefore subjects' causal knowledge about the inner workings of the technical system was enhanced. In the ACT-R model we encoded this strengthening phase as the context in which subjects acquired and used their causal knowledge. However, each time this initial strengthening phase was followed by a short phase with negative evidence (discrediting phase). In this phase participants questioned their causal knowledge and the confidence in their predictions decreased. Thus, in the experiment we changed the experimental stimuli from strengthening causal relations to discrediting the very same. Psychologically, this change might have appeared as a contextual change. Suddenly the working principle of the technical system changed. This change was not visible to the participants, except that their learned causal relations did not lead to successful prediction of the state of the output system. Again, psychologically participants might have represented this change as a step from one working principle to another and so considered a new context.

For theories of causal learning the presented model raises some questions. From their perspectives peoples judgments and confidence ratings are seen as the result of reasoning or data integration processes. Since our simple memory model modeled the decrease in participants' confidence ratings perfectly, these more complex standard views of seem questionable. As introduced in the beginning, standard theories assume much more than memory retrieval. There are much less approaches that assume, as we do that lower ratings in causal reasoning might reflect a reduced availability i.e. accessibility of memory information due to less activation. In our model this goes back to less relevance of memory information as soon as new contexts are considered. Thus 'deliberate' causal behavior can be explained simply based on memory activation.

Of course, further experiments should replicate the present work. Additionally, it should first be tested whether our contextual change (ACT-R) model holds in other situations of causal learning as well. Second, it has to be proven for more cases that confidence ratings can be drawn from activation. The non-linear transformation function that was used has to be tested. It assumes that the higher the activation the less are changes of that activation reflected in confidence ratings drawn from it, even if the base-level learning that generates these activations already shows this kind of non-linearity.

Our next step is to elaborate more on the devaluation effect and the role of memory in causal learning and reasoning. For example if this effect occurs also for more complex causal knowledge or if similar effects can be found in reaction time data too. The presented ACT-R model will have to prove his validity for those data as well.

References

Anderson, J. R. (2007) How Can the Human Mind Occur in the Physical Universe? New York: Oxford University Press.

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.

Block, R. A., & Reed, M. A. (1978). Remembered duration: Evidence for a contextual-change hypothesis. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 656-665.

Drewitz, U. & Thüring, M. (2009). Modeling the Confidence of Predictions: A Time BasedApproach.In A. Howes, D. Peebles, R. Cooper (Eds.), 9th International Conference on Cognitive Modeling – ICCM2009, Manchester, UK.

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

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. Annual Review of Psychology, 62, 451–482.

Hertwig, R., Herzog, S. M., Schooler, L. J. & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. Journal of Experimental Psychology - Learning, Memory, and Cognition, 34(5), 1191-1206.

Kahneman, D. (2003). "A perspective on judgment and choice: Mapping bounded rationality". *American Psychologist* 58 (9): 697–720.

Mehlhorn, K., Taatgen, N.A., Lebiere, C., Krems, J.F. (2011). Memory Activation and the Availability of Explanations in Sequential Diagnostic Reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37*, 1391-1411.

Perales, J. C., & Shanks, D. R. (2007). Models of covaria- tion -based causal judgment: A review and synthesis. Psychonomic Bulletin & Review, 14, 577-596.

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. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.

Schooler, L. J. & Hertwig, R. (2005). How Forgetting Aids Heuristic Inference. Psychological Review, 112(3), 610-628.

Acknowledgement:

We thank Niels Taatgen and Hedderik van Rijn who initially had the idea to implement the concept of contextual changes.