Modeling of expectations and surprise in ACT-R

Stefan Lindner (stefan.lindner@zmms.tu-berlin.de)
Research Training Group "prometei", TU Berlin, Marchstrasse 23
10587 Berlin, Germany

Nele Russwinkel (nele.russwinkel@tu-berlin.de)
Department of Cognitive Modelling in dynamic HMI, TU Berlin, Marchstrasse 23
10587 Berlin, Germany

Abstract

Expectations are essential to model well-learned-tasks in human-machine interaction. Our aim is to implement expectations and surprise in the cognitive architecture ACT-R. We present a reaction time experiment designed to inform the structure of expectations and discuss its implications for the modeling of expectations. We also discuss general constraints that limit possible models of expectations and surprise in ACT-R and present future modeling and experimental ideas.

Keywords: expectations; surprise; cognitive modeling; ACT-R; usability; applied research

Introduction

Expectations are essential to our everyday life. Without expectations we would be hopelessly lost in the complexity of our environment. They help us to perceive a manageable environment that we can understand and act in. If our expectations are disappointed it marks an important event for us and we experience the emotion of surprise.

Depending on our expectations we make different forecasts about the future state of our environment and about the future worlds that can result from our own actions. By coupling those beliefs about the world with our goals and preferences, expectations guide our behavior and enable us to function in the complex and dynamic environment surrounding us.

Of course expectations also play an important role in human-machine interaction (HMI) since especially in well learned tasks our behavior is guided by expectations. Intuitive design, for example, tries to conform to widely held expectations about functions and forms within a technical artifact (Hurtienne & Blessing 2007). Expectations also feature importantly in user experience and performance in system switches. For instance, someone switching from an old cell phone to a newer model might expect a somewhat similar design and implementation of the functions. If some user input is not followed by the expected device output the user might feel irritated or disoriented. Violating expectations concerning the new system might drastically undermine acceptance. In the general context of HMI it can also lead to orientation loss and lead to critical mistakes in safety-relevant procedures. Reaction capabilities can be reduced when an unexpected event (e.g. an alarm) occurs. For example, Russwinkel et.al. (2011) found a strong distortion in time perception when the task switched unexpectedly. In the worst case, an unexpected environment can lead to a complete lack of understanding and action.

In order to properly model not only human-machine interaction but also human behavior in general, it is therefore imperative to take into account expectations and their consequences for human actions and thought processes. Our goal is therefore to create models of expectation and surprise that are general and accurate enough to be used as a component in other complex modeling efforts describing human behavior.

In the context of human-machine interaction, modeling is especially useful in the development phase of a technical device. It enables the prediction and evaluation of different aspects of user performance at an early stage of development. Those evaluations can in turn inform product design guidelines that can be passed on to product designers. Another strong point of modeling in the context of HMI is the ability to continually and quickly test device conceptions after they have been altered slightly.

Our goal is to implement expectations and the violation of expectations, surprise, in the precise terms of a cognitive architecture. Unfortunately, existing theoretical models of expectation and surprise are not specific enough to decisively inform an exact implementation yet. Our approach therefore is to first research the buildup of simple expectations in experiments. The data from these experiments are then compared to different cognitive modeling approaches. This way we hope to find mechanisms that can reproduce the effects found in the experimental data.

We chose to implement our models in the cognitive architecture ACT-R. The advantages of ACT-R in general and for our specific goal will be discussed in a later section.

Models of expectation and surprise

Expectations are anticipatory mental states that help to make forecasts about the future state of the world. In contrast to simple predictions, expectations have a belief and a goal component (Falcone & Lorini, 2005). While the belief component of expectations comprises the factual understanding of the world, the goal component defines the

personal relevance of the expectation by declaring a present or future state of the world as desirable or undesirable.

Surprise is typically defined as a discrepancy between held beliefs about a state of the world and observations about the actual state of the world. Meyer et. al. (1997) define surprise as a discrepancy between an activated schema (i.e. a building block of the knowledge structure) and the perception of a schema-discrepant event. Having grown up in New York you might, for example, have built up the expectation that cabs are always painted yellow. When you travel to London for the first time you see a black cab and you are surprised. Surprise facilitates the update of previously held beliefs and schemata by directing attention to the schema-discrepant events. In our example you might modify your old schema to now include black and yellow cabs and act accordingly, e.g. when looking out for cabs in foreign countries.

Surprise can be used as an indirect tool to deduce previously held expectations. In contrast to the expectations themselves, surprise results in several measurable behavioral artefacts. Surprise usually leads to a delay in task reaction time (Schützwohl, 1998; Niepel et. al. 1994). It is associated with the P300-component of EEG-measurements (Duncan-Johnson & Donchin, 1977) and leads to the activation of specific facial muscles (Ekman & Friesen, 1975).

According to Falcone and Lorini, at least three types of surprise can be distinguished. The authors postulate that there is always an active set of expectations that we are focused on and aware of at any specific point in time (for a possible formalization of the active expectation set see Fagin & Halpern, 1987). One type of surprise, *mismatch based surprise*, is generated by a mismatch between this active set of expectations and sensor data, i.e. data that can be accessed directly by the senses at a specific moment. The two other types of the surprise (*passive prediction-based surprise* and *implausibility-based surprise*), on the other hand, involve the reconciliation of sensor data with a wider general knowledge about the world.

Since active expectations, especially those concerning the immediate task at hand can be limited, described and manipulated better in experiments by presenting appropriate (simple) tasks, we chose to first concentrate on these expectations and the resulting mismatch based surprise. As mentioned before, the surprise reaction can manifest itself in many objective indicators. Currently we focus on reaction time differences elicited by experimentally induced expectation violations. The general idea of the research approach is to implement theoretical approaches of expectations and surprise with ACT-R models and to validate them with experiments. Starting off with very simple experiments that limit task relevant expectations, our goal is to implement simple models in ACT-R that describe the formation, use and disappointment of expectations in these experiments. This way we hope to arrive at a theory of expectations in ACT-R that can later be used to model much more complex and realistic situations, e.g. the aforementioned human-machine interaction involving a system switch.

ACT-R

As a cognitive architecture we used ACT-R because ACT-R is especially suited for describing memory mechanisms and thus lends itself well to modeling expectations. It is supported by a big research community that can provide input and ideas. Improvements in the architecture can also prove fruitful for many other modeling efforts. Last but not least, ACT-R is open source, so it allows for the integration of a self-developed expectation component.

ACT-R is a cognitive architecture that enables the modeling of goal-driven behavior (Anderson et al., 2004). It assumes a modular structure of the brain, i.e. separate subsystems being responsible for different tasks of perception, cognition and motion. Modules include a declarative memory, motor, vision and intentional (goal) module. Modules interact via their interfaces called *buffers*. These buffers are manipulated by the means of *productions*.

Productions are rules that contain a conditional part (the LHS) that specifies at which combination of buffer states the rules fires and an action part (the RHS) that specifies resulting buffer changes. Models in ACT-R consist of these production rules and, once started, run continuously until no conditional part (the LHS) of any production matches the current combination of buffer states.

Modeling human-machine interaction in ACT-R offers a great range of advantages. ACT-R is grounded in an extensive body of research findings and is continually being expanded as our understanding of the mind increases. ACT-R is implemented in a programming environment (currently the ACT-R 6 environment; Bothell, 2007), allowing for the construction of complex models and the modeling of complex and extensive behavior. It thereby exceeds the scope and possibilities of most traditional theoretical models in both regards. Another advantage of the ACT-R architecture is the easy coupling of the model of a technical device with the cognitive model of the user.

Experiment

The idea of the experiment was to first build up a task-specific expectation and to subsequently disappoint it. Reaction times were recorded and served as the main indicator of expectation violations and surprise. Time and frequency of the expectation violations were manipulated in order to gain greater insight into expectation buildup and actualization.

The specific setup repeatedly presents two geometrical figures in order to build up the task-relevant expectation that only these two figures would appear during the remainder of the trial. The appearance of a third figure then marks an expectation-violating event. We therefore expect its appearance to result in a surprise reaction and prolonged reaction times (hypothesis 1).

In a reaction time experiment that influenced our experimental setup, Schützwohl (1998) presented black-text-on-white-background slides. Participants had to react to a binary stimulus on the slides. At some point in the experiment participants were surprised with the presentation of white-text-on-black-background slides. In the critical (surprise) trials, reaction times were prolonged compared to reaction times immediately before the critical trial. The later the critical trial occurred for the first time in the experiment, the more the reaction times were prolonged.

Since expectations and schemata should strengthen with the sample size of confirming events in our experiment as well, we expected the surprise strength, and with it, reaction times to increase monotonically as a function of the first appearance of the schema-discrepant event (i.e. the first triangle) in our experiment, too (hypothesis 2).

Also informed by the Schützwohl experiment and by schema theory (G. Mandler, 1984), we expected the return of reaction times to the baseline after immediately consecutive presentations of the rare geometrical form in the second consecutive trial of its presentation (hypothesis 3).

Methods

In a reaction time task 100 participants (48 female; mean age=32.5, SD=12.1) were presented basic geometrical figures. Participants were instructed to react to the figures by pressing a number key that correspond to the number of the figure's edges as fast as possible. They were not informed about which specific figures would appear. Upon completion, participants were given some sweets regardless of performance.

In all experimental conditions two figures (squares and pentagons) were presented frequently and in a pseudo-randomized order. A triangle was very rarely presented and constituted a critical trial. In all experimental conditions, at least 14 squares and pentagons were presented before the first appearance of a triangle.

Experimental Conditions: All experimental conditions are depicted in figure 1. All geometrical forms had the same color (blue) but triangles are highlighted here for sake of clarity.

In the *control condition* participants were confronted with only squares and pentagons. The condition served to establish baseline reaction times.

In *condition 1.1* participants were confronted with an early triangle (15^{th} trial), in *condition 1.2* with a late triangle (40^{th} trial). These two conditions allowed establishing whether an early expectation violation resulted in a different reaction time delay than a late expectation violation.

In *condition 2.1* a triangle was presented in both the 15^{th} and the 40^{th} trial. This condition aimed at establishing whether after an expectation violation participants would still show a prolonged reaction time when the initially unexpected event happened again after some delay (here after 24 trials without a triangle) .

In *condition 2.2* we presented triangles in three consecutive trials both early (15^{th} -17^{th} trial) and late (40^{th} - 42^{nd} trial) in the experiment. This condition served to establish how fast reaction times would recede to the base line after an initial expectation violation.

For all trials reaction times were recorded. Additionally, participants were asked to rate the strength of their surprise when a triangle was presented in the 15^{th} trial on a continual scale from 0 ("not surprising at all") to 10 ("very surprising").

Results

Table 1 depicts reaction times for trials 15-17 and 40-42 in the control condition and for the critical trials in all other conditions. Because mean reaction times in trials 10-14 differed between the conditions ($F=6.514$, $df=4;20;24$, $p<0.01$), those means are given for comparison purposes, as

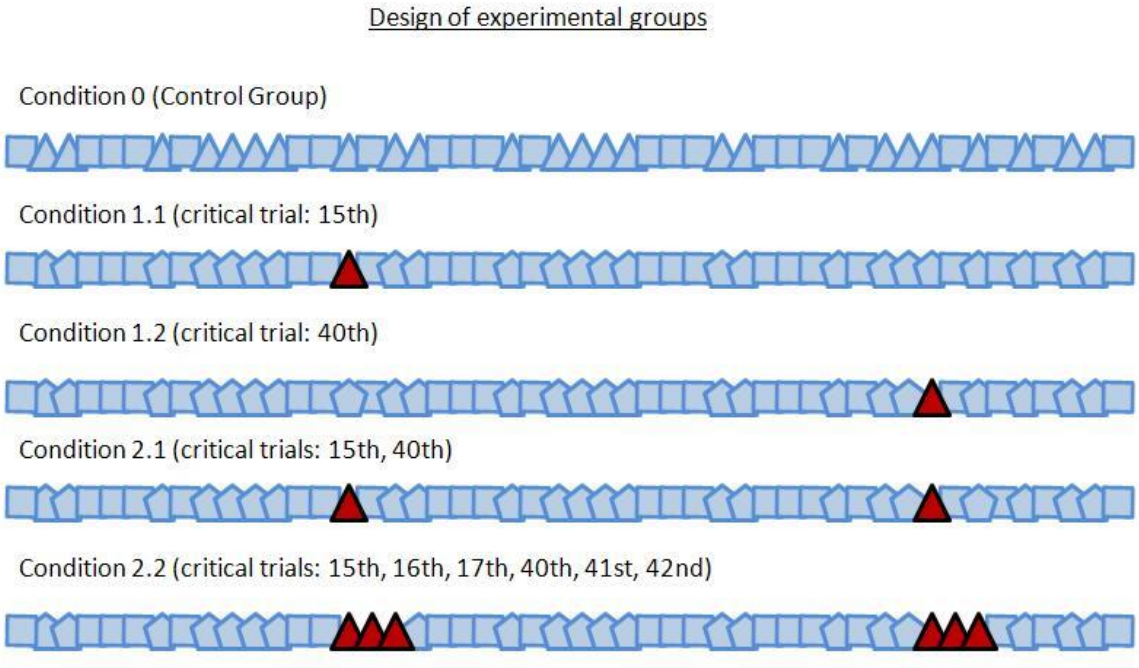

Figure 1: Experimental Setup

well. For the statistical analysis, reaction times were first normalized within each participant.

	Average RT in trial 10-14	15th trial	16th trial	17th trial	40th trial	41st trial	42nd trial
Condition 0 (control)	669 (75)	684 (83)	646 (83)	645 (86)	604 (86)	608 (80)	579 (66)
Condition 1.1	722 (124)	815 (96)					
Condition 1.2	766 (150)				889 (105)		
Condition 2.1	749 (138)	889 (124)			814 (144)		
Condition 2.2	793 (149)	886 (110)	702 (110)	718 (119)	806 (91)	656 (98)	641 (110)

Table 1: Reaction times [in ms]; (Standard deviations)

Reaction times in the critical trials (i.e. trials showing a triangle) were significantly longer then in the corresponding baseline trial, supporting hypothesis 1. Depending on the condition, reactions were prolonged by about 100-150 ms in the critical trials.

Somewhat unexpectedly reaction times in the early critical trial were not more prolonged then in later critical trials, lending no support to hypothesis 2.

In condition 2.1 and 2.2 we found that reaction times fell back to baseline immediately in critical trials following the initial critical trial as we had conjectured in hypothesis 3.

Participants also reported to be mildly surprised in a critical 15th trial, rating the presentation of a triangle with a mean score of 4.00 (n=43, SD= 2.41) on a scale ranging from 0 ("not surprising at all") to 10 ("very surprising").

Discussion

As expected, reaction times were prolonged in the critical trials. As we will also discuss in the modeling section, this could solely reflect longer retrieval times of relevant knowledge. In this specific case the delay could at least be partly due to the fact that it will take longer for participants to recall the fact that a triangle has 3 edges - as opposed to the matching rules for squares and pentagons – since this fact has not been needed before in this task. The prolonged reaction time could, however also reflect additional processing steps that take place after an expectation-discrepant or surprising event. Those processes could include the additional attentional focusing of the event, verification processes and evaluation processes (Meyer et. al. 1991).

In contrast, although we expected reaction times to increase monotonically as a function of the first appearance of the schema-discrepant event, our results failed to show this effect. A possible explanation could be a ceiling effect of schema strength in our specific experimental setup.

The most interesting and possibly most consequential experimental result though is the immediate return of reaction times in critical trials immediately following a critical trial. This result is in line with schema theory, which predicts a quick schema update in case of a schema-discrepant event (G. Mandler, 1984). As we will discuss later in the general modeling considerations, this result might also be a very important constraint on ACT-R models of expectations and surprise.

General Modeling Considerations

Before we describe and discuss the specific ACT-R models that we devised, we first quickly discuss general guidelines that model building should be constrained by:

1) Models of surprise and expectations should reflect the state of current research.

This can be achieved by using well grounded ACT-R mechanisms to model theoretical approaches of expectation/surprise. It can also be achieved by postulating new mechanisms in ACT-R that can account for surprise as well other reactions that involve arousal.

2) Models should be as simple as possible while still being functional.

This principle concerns both the complexity of the model (i.e. the productions used) as well as the extent with which traditional ACT-R mechanisms are replaced by direct manipulations by the modeler. Following Occam's razor, if two models describe a phenomenon equally well, we should always prefer the simpler model.

Modeling Approaches

We devised three models to describe the experiment and to predict experimental reaction times. One model (the "activation model") was a simple model that relied solely on the activation of declarative knowledge. The two other models (the "threshold model" and "schema update model") were modeled following a theoretical surprise model and contained a new chunk construction that we will call "expectation-chunk" in the following.

Simple activation model: In the case of the presented experiment one of most frugal models that should be considered is a model that relies solely on the activation of chunks that specify the number of edges for each basic geometrical form. Such a model was indeed considered in the modeling effort of the experiment. After the participant visually encodes the presented figure, a chunk is recalled from the declarative memory that specifies the number of edges of the figure. Afterwards the corresponding key is pressed and attention is focused on the next trial. Although this simple "activation model" predicted both relative and absolute trends in experimental reaction times best (compared to the other 2 models considered) it failed to account for two important sets of data points.

Critically, in the case of repeated presentation of the triangles in consecutive trials, it failed to predict the immediate return of reaction times to the baseline in the second critical trial. The model therefore does not seem to account for the fast schema update taking place after the

initial surprise. We judge it somewhat less important that the simple activation model also fails to capture the severely prolonged reaction times in the very first trials. Many distractions and expectations not immediately relevant to the task at hand (e.g. considerations about the general nature of the experiment) could have been relevant in the beginning of the experiment. Those factors may also have been very diverse in nature and may vary between subjects, a conjecture that is supported by the fact that reaction times show a large variance in the very first trials.

Threshold model and schema update model: Our modeling efforts also included two attempts to model surprise following the structural model of surprise by Meyer et. al. (1997). The most extensive model (the "threshold model") contains productions that mirror an expectation-discrepancy threshold that had to be crossed in order for a surprise reaction to occur. It also contains productions that try to mirror a schema-update mechanism. The somewhat less extensive of the two models (the "schema update model") did not contain the threshold-mirroring productions anymore but retained the schema-update mechanism.

The two more extensive models importantly also included the new concept of "expectation chunks", i.e. chunks that are called to the short-term working memory (the "imaginal" in the ACT-R architecture) in anticipation of upcoming events. In our case the expectation chunks already contained the correct figure - edge assignments, so in the case of a correct "expectation" the correct rules did not have to be retrieved after the observation but was already present and enabled a faster reaction.

Unfortunately, both of the extensive models greatly overestimated the absolute reaction times.

Our experimental results put considerable constraints on possible models and one of those constraints arrives from the fact that in the critical trials reaction times were only about 100-150 ms longer then in the noncritical trials. In the ACT-R architecture the firing of a production takes at least 50 ms (Anderson et. al. 2004). This restricts us to very few (2-3) additional productions that can model the expectation management and surprise processes in critical trials taking place before the key press.

The fact that the more extensive models failed to reflect real reaction times can largely be attributed to violating this constraint. Both models postulated a large number of productions and multiple chunk retrievals which time requirements far exceeded actual experimental reaction times.

Outlook to a new modeling approach on expectation modeling

One of the central assumptions underlying the past and current modeling efforts is the existence of expectation chunks specifying expected future conditional and unconditional changes in the environment. Retrieval of these expectation chunks prepare an action response and enable a faster reaction should the expected situation arise.

In the models presented the expectation chunk was present in the "imaginal" buffer when the geometrical figure appeared. Since buffers can only hold one chunk at a certain point in time, the postulation of the existence of expectation chunks almost inevitably also leads to the theory that only one expectation can be immediately used in an action relevant way should the expectation hold true. This does not exclude other expectations from playing a role, though. Depending on the specific ACT-R model, highly activated expectations can quickly be retrieved and become action relevant as well. In fact, the models that included the use of expectation chunks ("threshold model" and "schema update model") also strived on making other highly activated chunks relevant. In future models we plan to make further use of expectation chunks, but the resulting theoretical implications have to be verified in experiments. Modifications in the structure and the exact application of expectation chunks are also possible.

An additional assumption that can be implemented in future models is that expectation chunks that are relevant in surprise reactions receive an activation boost that exceeds the activation they would receive from a simple retrieval. This boost reflects the enhanced attention that the expectation-discrepant situation receives in a surprise reaction. In practical terms this activation boost is supposed to model the fast schema update mechanism that was found in experiment 1 and also in Schützwohl's (1998) experiment. In the threshold model and the schema update model we tried to model this activation boost in a drawn-out mechanism that included the repeated retrieval of the relevant expectation chunk. However, as already mentioned, time restraints in the experimental data make those mechanisms implausible.

The described activation manipulation, however, constitutes an override of the intended traditional ACT-R mechanisms and must therefore be justified. Overrides of this kind should be held to a minimum in general, since each override to some extend abandons the empirically well founded basis of ACT-R. On the flipside, new additions to ACT-R are a necessary part of continually expanding the scope and explanatory power of the cognitive architecture.

Outlook and Discussion

Overall, the presented experiment informed modeling of expectations in two main points. First, it set a time constraint for productions that model the reaction delay in an unexpected situation. Second, the immediate return of reaction times to the baseline in case of the repeated exposure to expectation-violating events suggests that a simple activation model cannot fully model the processes that take place in such situations.

Thus model approaches different from those employed here are needed to satisfactory model expectation violating-situations.

Different experimental approaches are also in our focus both to improve on the current setup and to encompass different aspects of expectations and surprise as well.

One possible experimental change could aim at improving the strength of the two basic components of expectations, goals and beliefs. In the current experiments, subjects have the experimentally induced goal to react quickly to a shown geometrical figure with a key press. The subjective importance of this goal might vary with the competitiveness of the subject or the immersion in the task. Since the subjects received a small flat reward upon completion, goal strength likely remained rather weak, though. A way to address this problem might be to reward subjects based on their performance. A possible framework that avoids having to set arbitrary reaction time limits would be a competition among the subjects.

The experiments were designed to induce belief formation by leaving open the specific geometric figures to come. The repeated presentation of only squares and pentagons should have been able to lead expectations in the intended way, but subjects had to fight a relatively strong prior that triangles would appear at some time during the experiment. Subjects in the control condition stated that they allotted a 66% probability to that event before the experiment had started. The appearance of a triangle might therefore have been a somewhat weak expectation violation. To better control for belief strength, but also to gain more insight into the formation of expectations, we already started experiments in which participants are prepared differently for the events to come, either with an implicit or explicit instruction or with a preparatory exercise.

Another problem that arises in the current setup is the focus on only one dependent variable, reaction time. Reaction times are currently the only means to judge our models, the underlying processes might be complex though. Several competing ACT-R models, each postulating very different mechanisms might all be able to explain current experimental reaction times. The situation therefore resembles an underspecified equation system. In order to better differentiate between possible models of expectation and expectation violation, we might therefore be advised to gather additional dependent measures. We are restricted in those measures, though, by the variables that can be predicted by ACT-R.

One possibility is the additional measurement of BOLD-levels (Logothetis, 2002) that can lead to better information about the activation of certain brain regions and, relevantly, the associated modules. Discerning which modules are active at which points of time during the task could help to greatly reduce possible ACT-R models. Another possibility would be to offer participants a wide choice of reasonable action choices within the experimental task. This would inevitably require a more complex experimental setup than is currently employed. The obvious choice, given the background of our modeling efforts would be an interaction with a technical device (e.g. a smart phone). In order to keep expectations controlled, the interaction would have to take place in a relatively contrived, artificial context. With the ACT-R models then having to predict action choices as well as reaction times, not only is the modeling space much more restricted but models also move much closer to the intended long-term goal of predicting behavior in human-machine interaction and other complex situations.

Acknowledgments

This work was sponsored by Deutsche Forschungsgemeinschaft (DFG Research Training Group Prospective Design of Human Technology Interaction, GRK 1013).

References

Anderson, J.R., Bothell, D., Byrne, M.D., Douglas, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111, 1036-106.

Bothel, D. (2007). ACT-R 6.0 reference manual. Working draft. From the ACT-R web site. <act-r.psy.cmu.edu/actr6/reference-manual.pdf>

Duncan-Johnson, C.C., & Donchin, E. (1977). On quantifying surprise: The variation in event-related potentials with subjective probability. *Psychophysiology*, 14, 456-467.

Ekman, P. & Friesen, W.V. (1975). *Unmasking the face*. Englewood Cliffs, NJ: Prentice Hall, Inc.

Falcone, R., Lorini, E., (2005). Modeling expectations in cognitive agents, in: *Proceedings of AAAI 2005 Fall Symposium-From Reactive to Anticipatory Cognitive Embodied Systems*, 114–121, Menlo Park, 2005. AAAI Press.

Hurtienne, J., Blessing, L. (2007). Design for Intuitive Use - Testing image schema theory for user interface design, in: *Proceedings of 16th International Conference on Engineering Design*, Paris, 2007.

Logothetis N.K. (2002). On the neural basis of the BOLD fMRI signal. *Philos. Trans. R. Soc. London Ser. B.* 357:1003–37

Mandler, G. (1984). *Mind and body*. New York: Norton.

Meyer, W.-U., Niepel, M., Rudolph U., & Schützwohl, A. (1991). An experimental analysis of surprise. *Cognition and Emotion*, 5, 295-311.

Meyer, W-U., Reisenzein, R., & Schützwohl, A. (1997). Toward a Process Analysis of Emotions: The case of surprise. *Motivation and Emotion*, 21(3), 251-274.

Niepel, M., Rudolph U., Schützwohl, A., & Meyer, W.-U. (1994). Temporal characteristics of the surprise reaction induced by schema-discrepant visual and auditory events. *Cognition and Emotion*, 8, 433-452.

Russwinkel, N., Urbas, L., & Thuering, M. (2011). Predicting temporal errors in complex task environments: A computational and experimental approach. *Cognitive Systems Research*, 12(3-4), 336-354, 2011.

Schützwohl, A., (1998). Surprise and schema strength. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24: 1182-1199.