Running Head: Memory Activation in Diagnostic Reasoning

Memory Activation and the Availability of Explanations in Sequential Diagnostic Reasoning

Katja Mehlhorn

Dept. of Psychology, University of Groningen, the Netherlands

Niels A. Taatgen

Dept. of Artificial Intelligence, University of Groningen, the Netherlands

Christian Lebiere

Dept. of Psychology, Carnegie Mellon University, USA

Josef F. Krems

Dept. of Psychology, Chemnitz University of Technology, Germany

Abstract

In the field of diagnostic reasoning, it has been argued that memory activation can provide the reasoner with a subset of possible explanations from memory that is highly adaptive for the task at hand. However, few studies have experimentally tested this assumption. Even less empirical and theoretical work has investigated how newly incoming observations affect the availability of explanations in memory over time. In this paper we present the results of two experiments in which we address these questions. While participants diagnosed sequentially presented medical symptoms, the availability of potential explanations in memory was measured with an implicit probe reaction time task. The results of the experiments were used to test four quantitative cognitive models. The models share the general assumption that observations can activate and inhibit explanations in memory. They vary with respect to how newly incoming observations affect the availability of explanations over time. The data of both experiments was predicted best by a model in which all observation in working memory have the same potential to activate explanations from long-term memory, and in which these observations do not decay. The results illustrate the power of memory activation processes, and show where additional deliberate reasoning strategies might come into play.

Keywords: Memory Activation; Hypothesis Generation; Sequential Diagnostic Reasoning; Abductive Reasoning; Cognitive Model

Memory Activation and the Availability of Explanations in Sequential Diagnostic Reasoning

A basic goal of human cognition is to explain and understand the events happening in the world. Whether it is in scientific discovery, medical diagnosis, software debugging, or social attribution, people try to find explanations on the basis of what they observe. The kind of reasoning underlying this task is often called abductive (Josephson & Josephson, 1996) or diagnostic reasoning (Kim & Keil, 2003) and described as highly complex. First, complexity arises from the large number of potential observations that can each have a large number of potential explanations. Take for example a physician, who is confronted with a patient's symptoms. Each of the symptoms has a number of possible alternative explanations and only the combination of symptoms allows for selecting a diagnosis. The task is further complicated by the fact that information often does not become available all at once, but only over time. Even if given all at once, observations might be perceived and understood only over time due to limited cognitive capacities. Thus, the ability to integrate newly incoming information over the course of the diagnosis process is important. A related factor is uncertainty. The physician can never be sure if all symptoms necessary to find the correct diagnosis were observed and whether all observed symptoms were caused by the current disease. Despite all these constraints, people often generate explanations with high speed and accuracy (Johnson & Krems, 2001).

Theories trying to understand diagnostic reasoning consistently make the distinction between, on the one hand, the generation of a potential set of explanations or hypotheses and, on the other hand, the evaluation of these explanations or hypotheses against potential alternatives. Often the evaluation of hypotheses is assumed to be performed in a second, deliberate reasoning stage after a first stage in which potential hypotheses are generated from memory (e.g. Evans, 2006; Kintsch, 1998; Thomas, Dougherty, Sprenger, & Harbison, 2008; Wang, Johnson, & Zhang, 2006a). For the deliberate stage of hypothesis evaluation, a number

of strategies that allow reasoners to deal with the complexity of the task have been investigated (cf. Johnson & Krems, 2001). However, a key aspect of diagnostic reasoning is that observations can be associated to a large number of possible explanations in memory (in fact, the number of potential explanations has been shown to be computationally intractable; Bylander, Allemang, Tanner, & Josephson, 1991). Generating and deliberately evaluating the complete set of explanations is therefore often impossible due to constraints set by cognitive capacity and time available for diagnosis (Dougherty & Hunter, 2003a, 2003b). Consequently, already during the generation of explanations from memory a selection amongst potential alternative hypotheses has to be made (Dougherty, Thomas, & Lange, 2010; Thomas, et al., 2008).

The goal of this paper is to more closely investigate how memory activation processes can provide the reasoner with such an adaptive selection. Specifically, we want to test how memory activation can help the reasoner to select amongst a large number of potential explanations and how this selection is affected by newly observed pieces of information over time. In the remainder of the introduction we first give a short overview of empirical findings on hypothesis generation and then we take a closer look at the theoretical background.

Empirical Findings on the Generation of Explanations

"Although the evaluation of prespecified hypotheses has been the subject of research for many years, relatively little research has been concerned with the initial generation of the to-be-judged hypotheses." (Thomas, et al., 2008, p. 158; see also: Weber, Böckenholt, Hilton, & Wallace, 1993). Existing empirical findings concerning hypothesis generation consistently show that reasoners generate only a subset of up to four possible hypotheses from memory (Barrows, Norman, Neufeld, & Feightner, 1982; Dougherty, Gettys, & Thomas, 1997; Dougherty & Hunter, 2003a; Elstein, Shulman, & Sprafka, 1978; Joseph & Patel, 1990; Mehle, 1982; Weber, et al., 1993). Whereas this small number of generated hypotheses seems

to contradict the large number of potential hypotheses, research has shown that the selection of hypotheses into the generated subset is highly adaptive. Out of all potential hypotheses, reasoners generate those hypotheses that have a high likelihood of being relevant as explanations in the current situation. Specifically, those hypotheses seem to be generated that (a) have a high *a priori* probability based on previous experiences (Dougherty, et al., 1997; Dougherty & Hunter, 2003a; Gettys, Pliske, Manning, & Casey, 1987; Sprenger & Dougherty, 2006; Weber, et al., 1993) and (b) that are most likely in the context of the current observations (Weber, et al., 1993).

Whereas the studies mentioned above say something about the outcome of the hypothesis generation process, they say little about the cognitive processes that yield this outcome (exceptions are Dougherty & Hunter, 2003a, and Dougherty & Sprenger, 2006, who showed that participants tended to generate those hypotheses that were most 'active' as defined by a strength manipulation in the learning phase). To test if memory activation can indeed help the reasoner to select explanations from memory, the availability of explanations has to be assessed as a function of the observed information. In previous experiments, the availability of explanations has been estimated using explicit measures. For example, Wang, Johnson and Zhang (2006b) asked their participants for explicit belief ratings after serially presented observations and Dougherty and Hunter (2003b) asked their participants for probability judgments of different explanations. However, such explicit measures have two major drawbacks. First, explicitly asking participants during the course of the task might influence the outcome of the task itself (cf. Hogarth & Einhorn, 1992). Second, although there have been efforts at clarifying this issue (Drewitz & Thüring, 2009; Thomas, et al., 2008), it is not clear how the implicit concept of availability in memory translates into explicit concepts like ratings and judgments. Furthermore, to investigate how the availability of explanations is affected by newly incoming observations, availability should be tracked over time. With few

exceptions (Baumann, Krems, & Ritter, 2010; Sprenger, 2007; Wang, et al., 2006b) this issue has received little attention in previous studies.

Methods used in diagnostic reasoning research range from protocol analysis of physicians explaining a patient's pathophysiology (Arocha, Wang, & Patel, 2005) to simple laboratory experiments where only a few pieces of evidence and few alternative hypotheses need to be considered (e.g., Wang, et al., 2006b). While the first method allows for high external validity of aspects like task complexity, the second method allows for high control of aspects like previous knowledge. For analyzing the subtle effects of memory activation it is essential to have an optimal trade-off between both.

In the experiments reported in this paper we attempt to address the issues discussed above by designing experiments (a) in which participants have to generate explanations in a diagnostic task that is more complex than in previously reported studies and (b) which at the same time are controlled enough to study memory effects. During this diagnostic reasoning task, we assess the availability of explanations not only at the end of a trial, but we also track the availability while new symptoms are observed. We do this with an implicit probe reaction time measure, rather than with an explicit measure of the explanations' availability. This should reduce potential effects of the measurement on the outcome of the task itself. Before we present the experiments in detail, we discuss the potential role of memory activation for the generation and evaluation of explanations.

Memory Activation and the Generation and Evaluation of Explanations

To understand the role of memory activation in diagnostic reasoning, it is necessary to consider how diagnostic knowledge is represented in memory (Arocha, et al., 2005). A large number of studies has shown that with increasing experience in a domain reasoners develop knowledge structures whose content reflects the structure of the environment (Anderson & Schooler, 1991; Gigerenzer, Hoffrage, & Kleinbölting, 1991). To illustrate this using our

earlier example, a physician will have a stronger memory representation of a diagnosis that has occurred frequently in the past, compared to a rare diagnosis. Similarly, the association between symptoms and their potential diagnoses in memory will increase with increasing experience of their co-occurrence. Given such a highly adapted knowledge structure, data extracted from the environment can serve as a cue for the retrieval of diagnostic hypotheses from long-term memory (Arocha & Patel, 1995; Ericsson & Kintsch, 1995; Kintsch, 1998; Thomas, et al., 2008). An observation's efficiency as retrieval cue will depend on how strongly it is linked to the explanation in memory; the stronger the link, the more activation will occur (Anderson, Bothell, Lebiere, & Matessa, 1998).

So far, we looked at the question of how observed information can serve as retrieval cue for one associated explanation from memory. However, a key aspect of diagnostic reasoning is that pieces of information are usually associated to a large number of possible explanations. Retrieving them all from memory is often impossible due to constraints set by cognitive capacity and time available for diagnosis. To understand diagnostic reasoning it is therefore not only necessary to understand how one potential explanation is retrieved from memory, but also how a selection is made among all the possible alternatives. For selecting explanations from a set of alternatives it is necessary to evaluate the alternatives in the set. A factor commonly linked to the evaluation of explanations is their coherence with the data. In his *Theory of Explanatory Coherence* Thagard (1989a, 1989b, 2000) showed how a set of potential explanations can be evaluated purely based on the coherence between the explanations and the observed data. In the computational implementation of this theory, *ECHO*, pieces of information are represented by interconnected nodes that, depending on their coherence to each other, spread activation or inhibition. The theory predicts that explanations most coherent with the observed data are most strongly available (because they receive a large amount of activation) and that explanations that are associated to only some of the observations have a lower availability (because they receive some inhibition). Applied

successfully to explain phenomena in various domains the theory has been described as a "computationally efficient approximation to probabilistic reasoning" (Thagard, 2000, p. 95). However, in its original implementation it is only used to model the integration of information given at a certain point of time.

An extension of Thagard's theory that can account for sequential information integration has been proposed by Wang at al. (2006b; see also Mehlhorn & Jahn, 2009). They assume that activation and inhibition spreading from new observations will add to the activation of observations that were observed before. Referring to work on memory retention, they propose that the impact of observations decays exponentially with the square root of time. Consequently, over time observations should increasingly lose their impact on memory activation. This assumption is in contrast to recent findings that suggest that information in working memory seems to be subject to very little decay (Berman, Jonides, & Lewis, 2009; Jonides, et al., 2008; Oberauer & Lewandowsky, 2008) or even no decay (Lewandowsky, Oberauer, & Brown, 2009). Thus, whereas constraint satisfaction seems to be a plausible mechanism for information integration at a certain point in time, the integration over time leaves open questions. Furthermore, the implementation of the theory into a connectionist network makes it difficult to asses how such a hypothesis evaluation mechanism would interact with the constraints set by other aspects of cognition, like perception, memory, and deliberate decision strategies.

A theory that takes into account the effect of limited cognitive resources on hypothesis generation and evaluation has recently been proposed by Thomas et al. (2008). In their HyGene model, diagnostic reasoning is described as a two-stage process, where a phase of automatic memory retrieval of hypotheses is followed by a phase of deliberate hypothesis evaluation. The memory retrieval stage itself consists of two parts. The first stage is a prototype extraction process, in which a memory trace is derived from episodic memory which "resembles those hypotheses that are most commonly (and strongly) associated with

the data" (Dougherty, et al., 2010, p. 308). In the second stage, this prototype is matched against known hypotheses in semantic memory. If sufficiently activated by the prototype, hypotheses from semantic memory are placed in working memory where they can be evaluated by deliberate reasoning processes. Although the authors stress the importance of understanding sequential information integration and discuss possible related questions, they do not present predictions for the sequential integration of information. Such predictions are complicated due to the assumptions of two distinct memory systems that are involved in hypothesis generation. Would, for example, new observations lead to the retrieval of different prototypes from memory? And if so, what would be the effects on the availability of hypotheses that were activated by previously retrieved prototypes?

Given the open questions presented above, we were interested in whether memory activation can indeed explain the generation and evaluation of explanations as found in an experimental setting. To answer this question, we extracted the most essential elements of the theories presented above and implemented them into a general cognitive architecture. For avoiding additional questions that might arise from understanding the interaction of episodic and semantic memory we focus on the effects on semantic memory. The basic assumption of the theories mentioned above is that each observation can affect the availability of explanations in memory. If an observation supports a particular explanation, the observation will spread activation to the explanation and will make it more available to the reasoner. If an observation does not support a particular explanation, the observation will spread inhibition to the explanation and make it less available to the reasoner[1]. If an observation is completely unrelated to an explanation, the explanation's activation will not be affected. Following the idea of Wang et al. (2006b), we assume that if several observations are currently in the focus of attention (that is, stored in working memory) they can serve as a sort of 'combined retrieval cue' for explanations in long-term memory.

Current paper

As mentioned above, not much progress has been made in understanding how exactly sequentially made observations will affect memory activation over time. To shed light at this question, we implemented four different cognitive models. These models all share the general assumptions about memory activation and inhibition as presented above, but they vary with regard to how strongly newly incoming observations affect the availability of explanations over time. In a first model, *Model-Current*, at each point in time only the most recent observation affects the availability of explanations. This model is designed to test whether the assumption that sequentially observed symptoms serve as combined retrieval cue is necessary, or whether the activation and inhibition spread by the current symptom alone can fit the activation curves found in the experiments. In the remaining three models the observations serve as combined retrieval cue and, thus, all affect the explanations' availability. The models vary with regard to how strong each observation is weighed. One of the models, *Model-Time*, will test the assumption that observations are weighed according to the times since they were observed as proposed by Wang et al. (2006b). As a decay of information in working memory has been questioned (Berman, et al., 2009; Jonides, et al., 2008; Lewandowsky, et al., 2009; Oberauer & Lewandowsky, 2008) we implemented two alternative models in which observed information does not decay. One model, *Model-Constant*, will test the assumption that observations are weighed according to the total amount of information that is currently held in working memory. This assumption arises from the idea that the total amount of activation that can be spread from working memory is a limited and constant amount that will be equally divided between the elements in working memory (Lovett, Daily, & Reder, 2000). The fourth model, *Model-Number* tests the assumption that all pieces of information that are currently stored in working memory are weighed equally, independent of the time since their observation and independent of the number of observations. Consequently, in this model, the

total amount of activation and inhibition spread into long-term memory will increase with the number of observed symptoms.

To test the models, we conducted two behavioral experiments. In the experiments, participants had to find diagnoses for sequentially presented series of medical symptoms. The knowledge necessary to solve this task consisted of a number of symptoms each of which was associated to a number of alternative explanations. Whereas the symptoms were real medical conditions, their association with the explanations was artificial to avoid possible effects of prior knowledge and for being able to fully balance the material. To be able to investigate the effects of memory activation on the generation of explanations we tried to minimize the role of deliberate hypothesis evaluation strategies in the task. Therefore, experimental trials were generated in a way such that in most trials to find the correct diagnosis it was sufficient to retrieve the one explanation from memory that was most coherent to the set of observed symptoms. Thus, whereas each of the serially presented symptoms had a number of possible explanations that should vary in their availability over the course of the trial, at the end of the trial the most active explanation would also be the correct diagnosis. We expected the activation of explanations in memory to depend upon the serially observed symptoms as described above; with supporting symptoms increasing an explanation's availability and non-supporting symptoms decreasing its availability. Activation was measured with a probe reaction task. The idea behind this task is based on lexical decision tasks where participants respond faster to a probe that is more highly activated in memory than to a probe of lower activation (e.g., Meyer & Schvaneveldt, 1971). We will now first describe the method and the data from Experiment 1. Then we describe the cognitive models in detail and present the model results. Subsequently, we present Experiment 2, compare its results to predictions of the models and discuss the implications of our findings.

Experiment 1

The goal of Experiment 1 was to test whether the availability of explanations over the course of diagnostic reasoning indeed depends upon the information observed over time. Therefore, we tracked the activation of three different kinds of memory elements during trials of a diagnostic reasoning task: (a) explanations that were supported by all the observed symptoms (compatible explanations), (b) explanations that were not supported by all the observed symptoms (incompatible explanations), and (c) explanations that were completely unrelated to the symptoms (foils). If the availability of explanations in memory depends on the observed symptoms as described above, we would expect symptoms to increase the activation of compatible explanations and to decrease the activation of incompatible explanations. The availability of foils should not be affected by the observed symptoms.

To introduce some uncertainty in the task, we varied the reliability of the symptoms presented in each trial. Whereas in 75% of the trials all symptoms reliably pointed towards the correct diagnosis (coherent trials), in 25% of the trials a misleading symptom was added (incoherent trials). Participants were not told whether a trial was coherent or incoherent.

Method

Participants

Twenty-three undergraduate students from the Chemnitz University of Technology took part in this experiment. Of those, one participant had to be excluded from analysis, because she did not reach the required performance in the training session. Twelve of the remaining 22 students were female. The mean age was 24.1 (SD = 6.8).

Tasks

Diagnosis task. Participants where told that the main task they had to solve was to diagnose hypothetical patients after a ‘chemical accident’. In each experimental trial, a set of

three to four symptoms was presented and the chemical that explained the combination of these symptoms had to be found (see Figure 1 for a sample trial). This task allowed us to assess overall performance in the trials.

Probe task. The second task to be solved in the experiment was a probe task. After one of the symptoms in each trial, a probe was presented. Participants had to decide as fast as possible whether the probe (e.g., T in Figure 1) was the name of one of the chemicals learned in the training session or not. Participants where told that the two tasks were not related to each other. This task allowed us to track the availability of explanations over the course of the diagnosis task.

--- insert Figure 1 about here ---

Material

Learning material. The materials that participants had to learn before the experiment consisted of nine different chemicals (see Table 1). Chemicals were named with single letters, which allowed us to construct balanced, artificial connections between symptoms and explanations about which participants would have no prior knowledge. Furthermore, using single letters as chemical names allowed us to use letters in the probe task, avoiding potential problems associated to the use of whole words (for example, individual differences in reading-speed and word frequency effects). The chemicals were grouped into the three artificial categories Landin, Amid, and Fenton. Participants were told that chemicals from the three categories differed in their state of aggregation: Landin chemicals, for example, being gasiform and affecting especially the respiratory system because they are inhaled. This organization of knowledge into a hierarchical structure was used to ease the learning of the material by allowing participants to connect it to their knowledge about biological workings of the human body. It reflects in a simplified form the hierarchical knowledge organization found in medical diagnosis (Arocha & Patel, 1995). Each chemical caused three to four

medical symptoms. Symptoms either had a relatively small number of two or three explanations (specific symptoms like cough) or a larger number of six explanations (unspecific symptoms like headache). This variance in the number of explanations was introduced because it is an important feature of real world diagnostic knowledge that increases the complexity of the task.

--- insert Table 1 about here ---

Experimental material. Coherent trials were generated by presenting the three or four symptoms caused by one of the chemicals. In those trials all symptoms pointed consistently toward the correct diagnosis. Incoherent trials were generated by inserting an additional misleading symptom into the symptoms of one of the three-symptom chemicals (see Table 2 for a coherent and an incoherent sample trial). Apart from this manipulation, the order in which symptoms were presented in each trial and the order of trials were randomly chosen for each participant. Each diagnosis occurred equally frequent during the experiment. Participants were told that, throughout the experiment, the second symptom of each trial might be misleading[2]. To keep them aware of this, the second symptom of each trial was printed in normal letters, whereas all other symptoms were printed in bold letters. Participants had no means of distinguishing coherent from incoherent trials until they observed the third symptom of the trial which was either consistent with the second symptoms (coherent trials) or not (incoherent trials).

To track the activation of explanations a probe was presented after one of the symptoms in each trial. Each probe was a single letter that was either a target probe (one of the names of the nine chemicals), or a foil (see Table 2 for examples of the different probe types). Target probes were either *compatible targets* or *incompatible targets*[3]. Compatible targets probed explanations that were supported by all the symptoms preceding the probe (except for the misleading symptom in incoherent trials). Incompatible targets probed

explanations that were not supported by all symptoms. The incompatible targets were chosen such that they were not supported by at least the first symptom of the trial. This allowed us to test the possible effect of inhibition beginning directly after the first symptom, where explanations that were supported by the symptom (compatible targets) could be compared to explanations that were not supported (incompatible targets). *Foils* were randomly sampled from nine letters that were not associated to any of the symptoms (see Table 2).

The Type of Probe (compatible target, incompatible target, or foil) and the Position of the Probe in the trial (after the first, second, third, or fourth symptom) were randomized over trials, with the constraints that (a) target probes and foils appeared equally often and (b) probes of each type appeared with equal frequency at all the positions. In 8.3% of the trials no probe was presented. Instead, after one of the symptoms of those trials, participants were asked to provide the set of diagnoses they currently had in mind. These trials merely were intended to prevent participants from expecting a probe in each trial, and were not analyzed.

--- insert Table 2 about here ---

Procedure

Each participant completed 5 sessions, which took part over a maximum of 10 days, with the first and second session on consecutive days.

Training session. The first session was a training session to ensure a high familiarity with the material and the task. It consisted of several blocks that were repeated until participants solved them with at least 80% accuracy. First, participants were presented with the cover story 'diagnose patient after chemical accident' and with the complete knowledge (see Table 1). After a paper-and-pencil exercise in which they could use the table to write down which chemicals were associated with each symptom, participants had to study each chemical category separately on the screen. They were asked to memorize and report the name of the category, the chemicals and their respective symptoms. When they could report

complete knowledge of the category at least once without error, they completed two more training blocks for that category. In the first block, sets of symptoms were displayed on the screen and participants had to enter the chemical that caused this set of symptoms. In the second block, symptoms were presented sequentially on the screen. After each symptom, participants were asked to enter all chemicals from the currently practiced category that could explain the symptoms seen so far.

After the training blocks for the single categories were completed, participants could again study the complete material (Table 1). They were then presented with four training blocks for the complete material. The first block was identical to the final one in the single category training, but now all categories were tested. The second block was used to familiarize participants with the concept of incoherent trials, that is, they learned that the second symptom of each trial might be misleading. In the third block the probe task was introduced. After an explanation of the task, participants were presented with probes and had to decide whether they were targets (chemicals) or foils. The last block consisted of trials identical to the trials in the experiment. Participants were sequentially presented with symptoms. After one of the symptoms they had to react to a probe, and after all symptoms had been presented, they were asked for their diagnosis. Depending on a participant's performance, this session lasted between 60 and 90 minutes.

Experimental sessions. The experimental phase was split into four sessions. Each session began with a short practice block to refresh the participants' knowledge of the material. Afterwards participants solved 96 diagnostic reasoning trials, of which 75% were coherent and 25% were incoherent. The completion of the experimental trials in each session took about 30 minutes. Each trial was started self-paced. The symptoms of the trial were presented sequentially in the middle of the screen for 2 seconds each, with a fixation cross presented for 1 second in between (see Figure 1). After one of the symptoms in each trial, either the probe or the question for the current set of explanations was presented. The probe

appeared in the form of a letter and participants had to indicate if the letter was the name of a chemical by pressing a button on a response box. At the end of each trial, participants were asked to enter their diagnosis on a standard keyboard. Participants were instructed to solve the diagnosis and the probe task as accurate and fast as possible. Reactions times for probes and diagnoses were recorded from the moment that the probe / question for diagnosis appeared on the screen. After each input participants received feedback about their response accuracy.

Results

Probe reactions

To test the activation of explanations during the diagnostic reasoning trials, reaction times of correct probe responses were analyzed in coherent and incoherent trials with correct diagnosis. Scores above and below three standard deviations from the mean within each condition were excluded from analysis, resulting in the elimination of 1.7% of the correct probe responses[4].

Coherent versus incoherent trials. To test if the reaction time patterns differed depending on whether the third and fourth symptom were consistent (coherent trials) or inconsistent (incoherent trials) with the second symptom, we conducted an ANOVA[5] with the factors Coherence (coherent vs. incoherent trial) and Type of Probe (compatible target, incompatible target, or foil). Symptoms before Probe (three vs. four) was used as numerical regressor variable. Neither the main effect of Coherence ($F(1,21) = 1.642$, $p = .214$, $\eta_p^2 = .073$), nor any of the interactions involving Coherence were significant (Coherence * Type of Probe: $F(2,42) = 2.776$, $p = .074$, $\eta_p^2 = .117$; Coherence * Symptoms before Probe: $F(1,21) < 1$; Coherence * Type of Probe * Symptoms before Probe: $F(2,42) < 1$). Consequently, for further analyses we collapsed the data over the factor Coherence.

Compatible versus Incompatible versus Foil. Figure 2a shows the reaction times of the different probe types over the course of the trials. Table 3 shows the results of the

ANOVAS performed to analyze this data. First, an ANOVA with the factor Type of Probe (compatible target, incompatible target, or foil) and the numerical regressor Symptoms before Probe (one, two, three, or four) confirmed a significant interaction. To check whether this interaction was indeed caused by different slopes of all probe types, we conducted additional ANOVAs for each pair of probe types. They confirmed significant interactions for each pair, except for the pair compatible-foil. For this pair, we additionally looked at the main effect of probe type, which showed to be significant, confirming that compatible probes are reacted to faster than foils. To test the course of availability over the course of the trial in more detail, we conducted additional simple effects analyses for each probe type. They confirm decreasing reaction times for compatible probes and foils. Incompatible probes did not vary over the course of the trial. Finally, simple effects analyses for symptoms before probe revealed significant differences between the probe types after each but the second symptom of the trial.

--- add Figure 2 about here ---

--- add Table 3 about here ---

Diagnoses

To assess participants' performance in the diagnosis task we measured diagnosis accuracy and diagnosis times at the end of each trial. For the analysis of diagnosis time, wrong diagnoses and diagnoses exceeding three standard deviations from the mean were excluded (resulting in an exclusion of 1.8% of the correct diagnoses). Diagnosis accuracy was equally high in coherent trials (*mean* = 95.5%; *SD* = 4.1) and in incoherent trials (*mean* = 95.5%; *SD* = 4.0), $t(21) < 1$. The equivalency between the conditions was supported by a Bayes Factor t-test, which showed clear evidence in favor of the null hypothesis, $BF=6.13^{6}$. This shows that the participants could solve the task well and, again, that there was no effect of a trial's coherence. Participants' time for entering correct diagnoses was fast overall, but

was significantly slower in coherent (*mean* = 795 ms; *SD* = 211) than in incoherent trials (*mean* = 496 ms; *SD* = 125), $t(21) = 8.612$, $p < .001$[7].

Discussion

The results of the probe reaction task in Experiment 1 support the assumption that the availability of explanations over the course of diagnostic reasoning depends on the observed symptoms. Compatible targets (explanations supported by all symptoms) were responded to faster than incompatible targets (explanations not supported by all symptoms) and foils (not related to any symptom). This is in line with the prediction that explanations in memory receive activation from symptoms that support them. Incompatible targets were not only responded to slower than compatible targets, but also slower than foils. This is in line with the prediction that symptoms inhibit explanations that they do not support.

An unexpected result of the probe reaction task was that not only the reaction times to compatible targets decreased over the course of the trial, but also those to foils. Foils were letters that did not name chemicals and were therefore not related to any of the symptoms. Given a pure memory activation account, these letters should not change in their level of activation over the course of the trial, as they receive no activation or inhibition from any of the observed symptoms. A possible reason for the unexpected reaction time decrease might lie in our methodology. By presenting the probes with equal frequency after one of the four symptoms, we might have caused participants to be increasingly prepared to respond to the probe towards the end of the trial. Such an increasing response preparedness can be described by a hazard function (Chechile, 2003) and is comparable to the foreperiod effect (Vallesi, Shallice, & Walsh, 2007). The foreperiod effect is "usually observed when a range of variable FPs [foreperiods] occur randomly and equiprobably, [and] consists of reaction times (RTs) decreasing as the FP increases" (Vallesi, et al., 2007, p.466). In our experiments, participants knew that after one of the symptoms in almost every trial a probe would appear. The position

of the probes' occurrence was randomly and equiprobably distributed over the trials. With each symptom that went without a following probe, the likelihood for a probe increased. Participants could thus prepare for the probe and react slightly faster to it later on in the trial. Consequently, it is likely that part of the increase in response times to all probe types is caused by an increasing response preparedness over the course of the trial.

The manipulation of the symptoms' coherence had neither an effect on the probe reaction times nor on the accuracy of diagnoses. As explained above, participants could determine the correct diagnosis in incoherent trials by remembering that the second symptom of each trial is potentially misleading. A very simple strategy to use this knowledge would be to simply ignore the second symptom of each trial. Whereas such a strategy would lead to good performance in the incoherent trials and in most coherent trials, it would lead to suboptimal performance in a small part of the coherent trials, where ignoring the second symptom does not allow for unambiguously identifying the correct diagnosis (this was the case in 15% of the coherent trials)[8]. Nevertheless, a closer look at the probe reaction data seems to support such a strategy. Whereas reaction times differ significantly between the different probe types after the first, third and fourth symptom, they do not differ after symptom two.

Whereas the probe reaction time patterns are in line with our predictions, the comparison between verbal hypotheses and empirical data is usually reduced to a qualitative *descriptive level*. To test if memory activation, combined with ignoring the misleading symptom and increasing response preparedness over the trial, can also quantitatively *explain* the data, we developed computational cognitive models of the task. The model entail (a) the assumptions about memory retrieval as described in the introduction, as well as (b) the strategy to ignore potentially misleading information and (c) the participants' increasing preparedness to respond over the trial[9].

Model

Model description

For reaching maximum comparability between the models, we implemented them all within one modeling framework, the cognitive architecture ACT-R (Anderson, et al., 2004). From all the variants of potential modeling accounts we chose ACT-R because it puts a strong emphasis on processes underlying memory activation (Anderson et al., 1998; Anderson & Schooler, 1991) and integrates these processes with general assumptions about human cognition. It accounts for both sub-symbolic and symbolic components of cognition and, therefore, allows for the implementation of automatic memory processes as well as deliberate reasoning strategies and their possible interaction. It has received empirical support and validation from a large number of studies in a variety of research areas (ranging from simple list memory tasks, Anderson, et al., 1998; to language acquisition, Taatgen & Anderson, 2002; see http://act-r.psy.cmu.edu/ for an extended list of publications). Furthermore, ACT-R allows for modeling of the complete task, as solved by the participant. Thereby, without requiring additional assumptions about how the model maps on the experiment, it produces results that are directly comparable to human data. This is possible because the ACT-R theory predicts not only the probability and latency of retrieving facts from declarative memory, but also the time taken to perceive a stimulus and give a response (e.g. by pressing a key).

Knowledge about facts is represented in the form of chunks in ACT-R's long-term memory, which is commonly referred to as declarative memory. Chunks can represent observations (e.g., medical symptoms), as well as their potential explanations (e.g., medical diagnoses). Access to the chunks depends on their activation in memory (Anderson, 2007; Lovett, et al., 2000). Only chunks whose activation exceeds a certain amount, the *retrieval threshold* τ, can be retrieved. The probability, p, that a chunk i will cross the retrieval threshold, τ, depends on its activation, A_i, as described by Equation 1

$$p = \frac{1}{1 + e^{\frac{\tau - A_i}{s}}}, \quad (1)$$

where s reflects the amount of noise added to the chunk's activation.

If a chunk is activated strongly enough to be retrieved, its activation, A_i, determines the time required for the retrieval. The more active the chunk, the faster it can be retrieved. The time it takes to retrieve chunk i is a negative exponential function of its activation, A_i, as shown in Equation 2:

$$Time = Fe^{-A_i}, \quad (2)$$

where F is a parameter scaling the latency of retrievals.

The idea behind the concept of a chunk's activation, A_i, is that the strength of activation reflects the likelihood (specifically, the log odds) of the chunk to be needed in the near future (Anderson & Schooler, 1991). This likelihood is determined by three factors: the chunk's usefulness in the past, B_i, its usefulness in the current context, S_i, and a random noise component, ε :

$$A_i = B_i + S_i + \varepsilon. \quad (3)$$

The chunk's usefulness in the past is reflected by the base-level activation, B_i. ACT-R predicts that the more often a chunk has been retrieved from memory and the more recent these retrievals were, the higher its activation. This prediction can explain empirical findings that show explanations with high base-rates of occurrence to be generated more often and earlier than explanations with low base-rates. Although the effects of an explanation's previous use are an interesting aspect of memory effects in diagnostic reasoning, they are not the focus of the current paper. Therefore, base levels were kept at a constant level in the model. This was plausible because participants received extensive training on the task (leading to a saturation effect) and all symptoms and explanations appeared equally often in the experiment.

The important factor for our research question is the second part of Equation 2: the chunk's usefulness in the current context, S_i. A chunk's usefulness in the current context reflects the likelihood that the chunk will be needed given the information currently available from the environment. In diagnostic reasoning, the current context is defined by the to-be-explained observations (e.g., the medical symptoms displayed by a patient; Arocha, et al., 2005; Johnson & Krems, 2001; Thomas, et al., 2008). ACT-R predicts that an explanation *i,* that is stored in long-term memory, receives activation, S_i, from each observation *j* that is currently stored in working memory[10] as described by Equation 4:

$$S_i = \sum_j W_j S_{ji}, \tag{4}$$

where the amount of spreading activation, S_i, is determined by the associative strength, S_{ji}, between explanation *i* and observation *j,* scaled by the amount of activation that can be spread from working memory, W_j. As we will describe in detail below, we manipulated this scaling parameter, W_j, to implement different ways of sequential information integration in the different models. The associative strength, S_{ji}, represents the extent to which observation *j* increases or reduces the likelihood that the explanation *i* is needed from memory. This relationship can be described by a log conditional probability ratio (Anderson & Lebiere, 1998):

$$S_{ji} = \log \frac{p(\text{observation}_j \mid \text{explanation}_i)}{p(\text{observation}_j \mid \text{not explanation}_i)}, \tag{5}$$

where the numerator describes the probability that observation *j* has been observed when explanation *i* is needed (i.e., is valid in this context) and the denominator describes the probability that *j* has been observed when *i* is not needed. Using an example, the equation describes the probability for observing the symptom cough while having the flu divided by the probability for observing cough while not having the flu. As the likelihood to observe cough is higher when having the flu is higher than when not having the flu, Equation 5 predicts a positive associative strength between cough and flu. In contrast, if an observation

(cough) does not support an explanation (pregnancy), the likelihood to observe cough when the patient is pregnant decreases. This results in a negative associative strength.

While Equation 5 provides a good estimate for associative strengths between chunks, their exact calculation is often computationally intractable (Anderson & Lebiere, 1998). Following ACT-R, we approximate positive associative strengths, S_{ji}, between chunks as:

$$S_{ji} = S - \ln(fan_{ji}), \qquad (6)$$

where S is a parameter for the maximum associative strength between chunks in memory and fan_{ji} is the number of chunks i that are positively associated to a chunk j. Following this equation, an observation that is associated with only few explanations (e.g., a medical symptom that is specific to a certain group of diseases) has a lower fan and therefore a higher associative strength to the explanations than an observation that is associated with many explanations (e.g., a medical symptom that is associated to a variety of diseases). While the associative strength between positively associated symptom-explanation pairs can be estimated as shown in Equation 6, the estimation of 'negative associations' is problematic. Depending on the certainty that is assumed in the task, the values for S_{ji} resulting from Equation 5 would lie somewhere between $-\infty$ (if it is absolutely certain that an explanation can be excluded from consideration when a certain observation is made) and 0 (if it is not know whether a certain observation and explanation can occur together). As ACT-R provides no solution for this issue, we treat negative associative strengths as a free parameter that we estimate from our empirical data.

Four different models of sequential information integration

To implement the different assumptions of how observations might affect the availability of explanations over time, we used the parameter W_j. This parameter scales the amount of activation and inhibition that each observed symptom can spread to long-term memory. For reaching maximum comparability between the models, we kept the total amount

of *W* after the fourth symptom at a constant level between the models[11]. Consequently, in all four models, the same amount of activation is spread from working memory after all symptoms have been observed. The models vary in how this activation is distributed amongst the symptoms and in how it varies over the course of the trial in the following way:

Model-Current. In the first model, at each point in the trial, only the most recently observed symptom spreads activation and inhibition to explanations in long-term memory. We implemented that by setting W_j for each but the current observation to zero. The current observation was scaled with value *W*.

Model-Time: In the second model, all observed symptoms spread activation and inhibition. As proposed by Wang et al. (2006b), the amount of activation spread by each of the symptoms depends on the time since the observation was made. The most recently observed symptom is weighed most strongly. Earlier observations are weighed with a decayed strength, with the strength decaying exponentially in the square root of time:

$$W_j = W_{j-i}(1-d)^{\sqrt{t}}. \quad (7)$$

Model-Constant: In the third model, all observed symptoms spread activation and inhibition as proposed by Lovett et al. (2000). The total amount of activation that can be spread from working memory has a constant value *W*. If several observations *j* are stored in working memory, they share this total activation. Consequently, the more symptoms are observed, the smaller is the impact of each of these symptoms:

$$W_j = W / n. \quad (8)$$

Model-Number: In the fourth model, the total amount of activation spread from working memory at a certain point in time depends on the number of observed symptoms. Each symptom can spread a fixed amount of activation, resulting in an increasing amount of spreading activation and inhibition with an increasing amount of observed symptoms. Consequently, in this model the amount of activation spread by each of the observations

neither depends on the time since the observation was made, nor on the amount of observations. Each symptom is scaled with the same value W_j.

Model procedure

All models follow the same procedure, with the only difference between the models being the setting of parameter W as described above. The model code can be downloaded from http://www.ai.rug.nl/~katja/. As for the participants in our experiments, the models observe sequentially presented medical symptoms, diagnose the chemical that caused these symptoms and react to the probe that is presented after one of the symptoms. The knowledge necessary to solve this task (see Table 1) is represented in the models' declarative memory and consists of two different types of facts, represented as chunks. The first type reflects the possible symptoms. The second type represents the letters that can be presented during the experiment (chemicals and foils) and their associated information. Each letter is represented by a chunk that holds the letter's name, the information stating whether it is a chemical or a foil, and, for chemicals, the associated symptoms[12].

When a symptom is presented on the screen, the model moves its attention to the symptom, reads it, and retrieves its meaning from declarative memory. The symptom is then stored in working memory. This process is repeated for each observed symptom so that, over the course of a trial, working memory is successively filled with the observed symptoms. Stored in working memory, symptoms automatically spread activation and inhibition to explanations in declarative memory as described by Equation 4. To simulate the strategy of ignoring the potentially misleading symptom, the second symptom observed in each trial is not stored in working memory. When the question for the final diagnosis is presented on the screen, the model retrieves that explanation from declarative memory that receives the most activation from the symptoms in working memory and enters the respective letter. The letter representing the correct explanation is most strongly associated to the observed symptoms.

However, as described above, the different models vary in how the associative strength between the symptoms and their explanations are weighed. In *Model-Current*, only the current symptom spreads activation. Thus, at the point of diagnosis, only the last of the observed symptoms affects activation of explanations in memory. In the remaining models all observed symptoms spread activation at the point of diagnosis. In *Model-Time*, the strength of activation depends on the time since an observation was made. Consequently, even though all observations affect explanations' availability in memory, availability is most strongly affected by newer observations. In *Model-Constant* and *Model-Number*, at the point of diagnosis, each symptom is weighed equally strong. As the letter representing the correct explanation is most coherent with the symptoms, it obtains the highest amount of spreading activation and is the one most likely to be retrieved. However, as shown in Equation 3, due to random noise also in these models it can happen that an alternative explanation receives more activation and is incorrectly entered as diagnosis.

When a probe is presented, the models move their attention to the probe and retrieve the chunk representing the probe letter. If that letter is stored as a chemical, the models respond 'yes', if it is stored as a foil, the models respond 'no'. As described by Equation 2, the speed by which a chunk can be retrieved depends on its activation. The more spreading activation the chunk receives from the symptoms in working memory, the higher it will be activated and the faster the retrieval. Thus, as in human participants, the time the models need to respond to a probe can be used as a measure of the activation of explanations in memory. To simulate the participants' increasing response preparedness over the trial, the models retrieve expectations about whether the upcoming stimulus is a symptom or a probe. If the retrieved expectation is met by the presented stimulus, the stimulus is processed as explained above. If the expectation is violated, the models need to make a change to their expectation before they can process the stimulus. This change in expectation costs 50 ms. The later in the

trial the probe is presented, the higher the chance that it is expected by the models and that no time-costly expectation-changes have to be made[13].

Results and Discussion of the Models

The models were run for each participant on the trials that this participant had solved. As described above, the four different models varied in their setting of the values for the parameter, W_j, that weighs the strength of observations *j* in working memory. All other parameters were kept constant between the models. To fit the models, we estimated the speed and stochasticity of memory retrievals, the base-level activation of facts in memory and the amount of spreading activation from symptoms to explanations[14]. All other parameters were kept at the default values of ACT-R 6.0 (Anderson, 2007).

Following the analysis of the human data, we collapsed the models' data over the factor coherence. The resulting reaction times to the probes are shown in Figure 2b. Fits for the probe reaction times and the diagnostic performance reached by each model are shown in Table 4. All models produce the basic result that, overall, compatible probes are reacted to fastest. This happens, because in all models compatible probes receive more activation from the observed symptoms than all other probe types. Incompatible probes are in all models slower than or at about the same level as foils. This happens, because in all models incompatible probes receive inhibition as well as activation from the observed symptoms. The reaction times to foils over the course of the trial are identical in all models, because these reaction times are not affected by spreading activation. As in the human data, they decrease over the trial. In the models this decrease is solely caused by the varying expectations about upcoming stimuli, suggesting that part of the decrease of reaction times to all probes was indeed caused by an increasing preparedness to respond. All models produce comparable diagnosis times. The models differ in the course of activation for compatible and incompatible probes and in the accuracy of their diagnoses in the following way:

Model-Current. Merely using the current symptom at each point in time, the model produces a surprisingly good fit to the probe reaction pattern. The model produces no difference between probe types after the second symptom, because no activation and inhibition is spread to long-term memory at this point. After all other symptoms, reaction times for compatible probes are faster than foils because compatible probes receive activation from the current symptom. However, contrary to the human data, reaction times to compatible targets do not increase over the course of the trial. Incompatible probes are slower than foils, with a decrease of reaction times over the course of the trial. This happens, because incompatible explanations are explanations that are incompatible to at least the first symptom of the trial. Consequently, incompatible probes always receive inhibition from the first symptoms and they can receive inhibition, as well as activation from the later symptoms. The model has a poor diagnostic performance, which is not surprising, as only the last symptom of the trial affects activation of explanations at the point of diagnosis.

Model-Time. Letting the impact of observed symptoms decay over time, the model produces a good fit to the empirical probe reaction data. After the second symptom the difference between probe types is smallest, because at this point in the trial, only the decayed activation and inhibition of the first symptom affect explanations' availability. After all other symptoms, reaction times to compatible probes are faster and decrease over the course of the trial as the amount of spreading activation increases with each observed symptom. However, this decrease is much less pronounced than in the human data. Reaction times to incompatible probes also decrease, because the later in the trial, the higher the chance that incompatible probes not only receive inhibition but also activation from the observed symptoms. The model produces correct diagnoses in about half of the trials, because symptoms that are presented late in the trial have an over proportional impact on explanations' availability.

Model-Constant. Letting the observations at each point in time share a constant amount of total working-memory activation, also this model produces a good overall fit.

However, also here the visual inspection of the time course of explanations activation shows some deviations from the human data. In the model, at each point in time a constant amount of activation is spread from working memory. Consequently, compatible explanations stay at a constant level over the course of the trial (with a slight decrease caused by increasing response preparedness over the trial). Incompatible explanations stay at a constant and relatively high level of reaction times between the first and the second symptom, and than decrease considerably. The model produces a high proportion of correct diagnosis, which is only slightly lower than in the empirical data.

Model-Number. Increasing the amount of spreading activation and inhibition with each observed symptom, the model produces the best overall fit to the human data. As in the human data, reaction times to compatible probes do not change from the first to the second symptom, and decrease afterwards. This happens, because compatible probes receive an increasing amount of activation with each but the second symptom. Incompatible probes slightly decrease over the course of the trial as they receive inhibition as well as activation. As *Model-Constant*, the model does not reproduce the dip in reaction times to incompatible probes after the second symptom. The model produces the same proportion of correct diagnoses as *Model-Constant*, because after the last symptom of the trial they are identical due to the setting of the total amount of parameter *W* at this point.

--- insert Table 4 about here ---

Summarizing, all models produce the overall pattern of probe response times as found in the human data. The models vary in how well they fit details of activation levels over the course of the trials. Only *Model-Constant* and *Model-Number* are able to produce a high diagnostic performance, because they weigh all symptoms equally strong at the point of diagnosis. However, even these models underpredict the diagnosis accuracy as well as the diagnosis times found in the human data. This underprediction is caused by the fact that in part of the consistent trials ignoring the second symptom does not allow for finding a correct

diagnosis. Whereas, as discussed earlier, participants might try to remember the second symptom once they realize that they cannot distinguish between explanations otherwise, the models do not have such knowledge. Simply relying on memory activation they have no means to correctly distinguish between alternatives if they receive an equal amount of activation from the observed symptoms. This result is a good illustration of the importance for automatic memory activation to interact with deliberate reasoning. Whereas in most experimental trials it was sufficient to enter the diagnosis suggested by memory activation, in coherent trials where ignoring the second symptom lead to equal activation of alternatives, participants most likely used additional deliberate reasoning processes to find the correct explanation.

In the experiment participants had to diagnose coherent and incoherent sets of symptoms, because we wanted to add uncertainty to the task and because we were interested to see what happens in cases where memory activation alone might not be sufficient to find the correct explanation. As the empirical and model data for diagnoses and probe reactions suggest, participants dealt with that challenge by simply ignoring the potentially misleading symptom. They did so, although they were told to use all the presented symptoms for their diagnosis, they were trained to do so in the practice session, the information was only misleading in 25% of the trials, and ignoring the second symptom reduced diagnosis performance in 15% of the consistent trials. As suggested by the probe reaction data and the models, using this strategy was highly adaptive, because it allowed for finding the correct diagnosis by simply relying on memory activation in the vast majority of the trials.

Experiment 2

Experiment 2 had three main goals. First, we wanted to test the reliability of the key findings from Experiment 1 with an experimental setup that allowed us more control over participants' strategies. Therefore, symptoms in this experiment always consistently pointed

towards the correct diagnosis. During trials we again tracked the activation of compatible explanations (supported by all symptoms) and incompatible explanations (not supported by at least the first symptom) and foils (not related to the symptoms). Second, we wanted to investigate in more detail the availability of explanations that are associated to only part of the symptoms observed in the trials. Therefore, in this experiment we tracked the availability of an additional group of explanations: rejected explanations. These explanations are explanations that support the initial symptoms of a trial, but are not supported by symptoms presented later on in the sequence. Consequently, they have to be rejected from the set of potential explanations at some point in the trial. Being able to inhibit such no-longer-compatible explanations has been described as one of the crucial aspects for diagnostic performance (Dougherty & Sprenger, 2006). To assess the activation of rejected explanations over the course of the task, we compared the activation of explanations that were (a) rejected at different points in the trial and (b) measured at different time spans after rejection. Third, we wanted to test how well the different models generalize to a new data set.

Method

Participants

Twenty-nine undergraduate students from the Chemnitz University of Technology that did not participate in Experiment 1 took part in this experiment. Three of them had to be excluded from data analysis, as they did not reach the required performance in the training phase. The resulting 16 female and 10 male participants had a mean age of 22.8 (SD = 3.6).

Material

Training material. The material that participants had to acquire in the training phase (Table 5) was a slightly modified version of the material from Experiment 1. Again, chemicals were grouped into categories and caused three or four symptoms. Whereas in

Experiment 1 symptoms were caused by the chemicals of either one, two, or all three categories, symptoms in this experiment were either caused by chemicals of one category (specific symptoms like cough) or by chemicals of all three categories (unspecific symptoms like headache).

--- insert Table 5 about here ---

Experimental material. In the experimental phase participants solved trials that were comparable to the coherent trials of Experiment 1 (see Table 6 for a sample trial). The only difference was that now also rejected explanations were probed. These explanations varied in the point of their rejection during the trial and in the number of symptoms presented between the rejection and the respective probe. This manipulation resulted in three different types of rejected target probes: 'rejected-after-2' that could be presented after the second, third, or fourth symptom; 'rejected-after-3' that could be presented after the third or fourth symptom; and 'rejected-after-4' that could only be presented after the fourth symptom. This allowed us not only to investigate the course of an explanation's activation after its rejection, but also the potential effect of when it is rejected in the trial. To prevent participants from expecting a probe in each trial, in 14% of the trials no probe, but the question for the current diagnosis, was presented after one of the symptoms. Again, these trials were not analyzed.

--- insert Table 6 about here ---

Procedure

The experiment consisted of one training session and two experimental sessions. In both experimental sessions participants solved 170 diagnostic reasoning trials, with a five-minute break after half of the trials were completed. Except for this, the procedure was identical to Experiment 1.

The Models

To generate predictions for the data of this experiment we used the models as described above, with the only change that the models now did not ignore the second symptom of the trial. Except for the total amount of memory activation that was increased to reflect the higher number of observed symptoms in the trial, none of the parameters of the model were changed[15].

Results

Probe reaction

Reaction times of correct probe responses were analyzed in trials with correct final diagnosis. Scores above and below three standard deviations from the mean within each condition were excluded from data analysis, resulting in the elimination of 2.0 % of the correct probe responses. The reaction times to all Types of Probes are presented in Figure 3a. Due to the incomplete design, analyzing the data with standard analyses is difficult. Here we present analyzes for three subsets of the data that are most interesting to test our predictions. Subsequently we present the model fits, which cover the complete dataset.

Compatible versus Incompatible versus Foil. First, we tested whether our results for compatible and incompatible target probes and foils could be replicated. Therefore, we did the same analyses as in Experiment 1; detailed results of the corresponding ANOVAS are shown in Table 7. An ANOVA with the factor Type of Probe (compatible target, incompatible target, or foil) and the numerical regressor Symptoms before Probe (one, two, three, or four) confirmed a significant interaction. To check whether this interaction was indeed caused by different slopes of all probe types, we conducted additional ANOVAs for each pair of probe types. As in Experiment 1, they confirmed significant interactions for each pair, except for the pair compatible-foil. For this pair, we additionally looked at the main effect, which again showed to be significant, confirming that compatible probes are reacted to faster than foils. To

test the course of availability over the course of the trial in more detail, we conducted additional simple effects analyses for each probe type. They show that, for all probe types, reaction times decrease over the course of the trial. Finally, simple effects analyses for the symptoms before probe revealed significant differences between the probe types after each but the second symptom of the trial.

--- add Figure 3 about here ---

--- add Table 7 about here ---

Compatible versus Incompatible versus Rejected-after-2 versus Foil. To test how the activation of rejected explanations changes with time after their rejection, we analyzed the course of activation of explanations that were rejected after the second symptom. Detailed results of the corresponding ANOVAS are shown in Table 8. An ANOVA with the factor Type of Probe (compatible, incompatible, rejected-after-2, and foil) and the numerical regressor Symptoms before Probe (two, three, or four) showed no overall interaction, but a significant main effect of Type of Probe. To compare rejected-after-2 targets to each of the other probe types, we conducted additional pair-wise ANOVAS. They reveal that rejected-after-2 targets interact with compatible targets, but do not interact with or differ from incompatible targets and foils. To test rejected-after-2 targets' course of availability over the course of the trial, we conducted a simple effect ANOVA. It shows that also reaction times for these targets decrease over the course of the trial. Finally, simple effects analyses for reactions after two, three and four symptoms revealed significant differences between the probe types after the third and fourth symptom.

--- add Table 8 about here ---

Time since rejection. The analysis of rejected-after-2 targets that is reported above sheds some light at the course of explanations' activation after rejection. However, a potential problem with this analysis is that it confounds the time since rejection and the time of measurement. Systematic effects of the time of measurement (e.g., the foreperiod effect or the

number of compatible explanations at the point of testing) might thereby drown out the effects of the time since an explanation's rejection. Therefore, we conducted an additional analysis in which we compared the different types of rejected targets (rejected-after-2, rejected-after-3, and rejected-after-4) when tested after the fourth symptom. An ANOVA with the factor Probe Type (compatible, incompatible, rejected-after-2, rejected-after-3, and rejected-after-4) confirmed that after the fourth symptom, reaction times differed significantly between the Probe Types, $F(5, 125) = 5.085$, $p < .001$, $\eta_p^2 = .169$. Holm-corrected pair-wise comparisons showed that reactions to compatible targets were faster than reactions to all other probes ($p < .04$), except for probes rejected after the fourth symptom ($p = .172$). No other difference reached significance. This confirms the prediction that explanations supported by all symptoms receive most activation and suggests that the activation of rejected targets indeed differs depending on the time since rejection.

Diagnoses

Again, we assessed accuracy and time for entering the diagnoses at the end of each trial. For the analysis of diagnosis times, wrong diagnoses and diagnoses above and below three standard deviations from the mean were excluded (resulting in an exclusion of 2.5% of correct diagnoses). The high diagnosis accuracy (95.9%; SD = 3.9) and short time for entering correct diagnoses (574 ms; SD = 264) show that participants could solve the diagnosis task with high performance.

Models

Model predictions for the probe reaction times are presented in Figure 3b. The associated fits and the diagnostic performance reached by each model are shown in Table 4. The model that also produced the best fit in Experiment 1, *Model-Number,* generalizes best to the probe reaction data of Experiment 2. A visual inspection of the model predictions shows that this model predicts the time course of compatible and incompatible probes very well and

better than the other three models. For rejected probes the picture is less clear. *Model-Constant* and *Model-Number* make almost identical predictions for rejected probes. Whereas these predictions are very good for rejected-after-4 probes, *Model-Time* seems to predict the time course of rejected-after-2 and rejected-after-3 probes better. However, in interpreting these results, it should be kept in mind that all predictions of the best-fitting model, *Model-Number* are within the standard errors of the empirical data. Again, only *Model-Constant* and *Model-Number* are able to produce the high diagnostic accuracy as found in the empirical data.

Discussion

Experiment 2 had three main goals: (a) to replicate the findings about the availability of compatible and incompatible explanations and foils in a more controlled setup, (b) to allow a closer evaluation of the availability of rejected explanations, and (c) to test how well the models generalize to a new dataset. We were able to replicate the results for compatible and incompatible explanations. The inspection of rejected probes suggests some difference between these probes, depending on the time since their rejection. The model comparison reveals large differences in generalizability of the models. *Model-Number* predicts the probe reaction data time and the diagnostic performance well, whereas the remaining models show clear deviations from the data. *Model-Number* is able to predict the effects for compatible and incompatible targets and foils. More interestingly, it is also able to approximate the pattern of the different types of rejected targets. The explanations rejected at different points in time had not been probed in Experiment 1 and therefore it was not self evident that any of the models would be able to predict them.

Given that the parameters of the models were fit to Experiment 1 and not adjusted to the data of this experiment, also the best fitting model, *Model-Number*, reaches a lower fit in Experiment 2 than in Experiment 1. This is not surprising, as reaction times in Experiment 2

decreased more strongly than reaction times in Experiment 1. Reasons might not only be found in differences between the samples, but also in differences between the tasks of the two experiments. In Experiment 1, participants had to keep in mind that symptoms might potentially be misleading and therefore that the current explanation might have to be changed during the trial. In Experiment 2, no such uncertainty existed and therefore participants could allocate more resources to the probe task. By adjusting parameters characterizing the sample (e.g., duration of memory retrievals) and the task (e.g., how strong response preparedness increases over the trial), the model could be fit to produce reaction times closer to those of the humans. In the current paper we decided to forgo this adjustment, because we were interested to see how well the model generalizes to a new data set (see Böhm & Mehlhorn, 2009, for earlier versions of the models that were fit to part of this dataset). The fact that without parameter adjustment *Model-Number* was able to predict the major effects found in the human data lends additional support to this model, as the ability of a model to generalize to a new data set, without any further parameter adjustments, has been described as an important standard by which models should be evaluated (Marewski & Olsson, 2009; Pitt, Myung, & Zhang, 2002; Roberts & Pashler, 2000).

General Discussion

In diagnostic reasoning, reasoners have to generate and evaluate possible explanations for data observed from the environment. Whereas the number of potential explanations is often large, reasoners usually only generate and deliberately evaluate a small subset of explanations. Empirical research has shown that the selection of explanations into the generated subset seems to be highly adaptive to previous experience and the current reasoning context (Dougherty, et al., 1997; Dougherty & Hunter, 2003a; Gettys, et al., 1987; Sprenger & Dougherty, 2006; Weber, et al., 1993). However, although the idea that currently available observations affect the generation of explanations from memory seems obvious, few studies

have experimentally tested this assumption. Even less work has investigated how newly incoming observations affect the availability of explanations over time. The goal of this paper is to more closely investigate how automatic memory processes can provide the reasoner with an adaptive selection from memory over time. We report the results of two behavioral experiments that were designed to overcome potential problems of earlier studies. The results of the experiments are compared to predictions of four cognitive models. Implemented in the cognitive architecture ACT-R, these models test hypotheses about how sequentially observed information might affect the availability of explanations in memory over time.

In both experiments participants diagnosed quickly and with high accuracy. Whereas all models diagnosed equally fast, only the models that weighed each observation equally strong at the point of diagnosis (*Model-Constant* and *Model-Number*) were able to replicate the high diagnosis accuracy. The models reached this performance by merely relying on spreading activation between symptoms and explanations, suggesting that, given sufficient knowledge, memory activation can indeed provide the reasoner with a highly adaptive selection of explanations from memory. The models' underprediction of diagnosis performance in trials of Experiment 1 where memory activation alone was not sufficient to find the correct diagnosis shows where deliberate reasoning processes might come into play.

The probe reaction task proved to be a useful measure for the availability of different explanations over the course of the reasoning task. Whereas for the participants the probe task seemed unrelated to the diagnosis task, reaction times to probes of different explanations varied, as predicted, as a function of the observed symptoms over time. All models were able to reproduce the overall activation differences between explanations found in the human data. This lends support to the basic assumption of spreading activation and inhibition as it was implemented in all models. The models differed in their ability to reproduce the courses of explanations' activation over time. In Experiment 1, all models reach a high overall fit, with varying success in fitting details of the activation curves. Furthermore, all models, but *Model-*

Constant reflect the ignoring of the second symptom in their curves. The generalization test of Experiment 2 shows that *Model-Number* generalizes best to the new dataset. The success of this model suggests that the impact of observations on memory activation might depend neither on the time since an observation was made, nor on the amount of observations. Rather, the results suggest that all observations that are stored in working memory seem to be weighed equally at each point in time until an explanation is found.

Generalizing to Real World Diagnostic Reasoning

To allow for experimental control that was necessary to test our assumptions about memory activation, the experiments and models in this paper present a simplified version of diagnostic reasoning. In real world diagnostic reasoning the task characteristics, the memory representation and the reasoning strategies will often be more complex. This increased complexity raises a number of issues, which we will briefly discuss here.

An important issue for understanding real world diagnostic reasoning is the interaction of automatic processes as investigated here with more deliberate reasoning strategies. Our models assume a very simple strategy: observed symptoms are successively stored in working memory and, when asked for the diagnosis, the explanation that receives the most activation from the observed symptoms is retrieved from memory. Obviously, such a simple strategy oversimplifies diagnostic reasoning. Whereas we chose to implement such a simple strategy to test different assumptions about automatic memory activation processes over time, it is very likely that people use additional deliberate strategies. People probably start to retrieve possible explanations early on in the reasoning process (see e.g., Just & Carpenter, 1987, for evidence that people interpret evidence as soon it becomes available). Thus, presumably, not only are the sequentially acquired observations stored in working memory, but also potential explanations that have been retrieved from long-term memory. Such an additional strategy of retrieving explanations earlier in the reasoning process might explain some of the deviation

between our probe data and the model predictions. For example, all models underpredicted the decrease of the slope of reaction times for compatible targets in both experiments. If the reasoner additionally would retrieve candidate explanations and store them in working memory, these explanations would be available at low time cost. Therefore, mean reaction times to compatible targets would decrease over the course of the trial to a stronger extent than predicted by our pure activation based models.

The question about reasoning strategies is closely linked to another important question for understanding real world diagnostic reasoning. How do people represent the sequentially observed data and the generated explanations in working memory? As discussed above, for the sake of simplicity, in our models only observations are stored in memory. Storing observations is not implausible, as it has been found that not yet explained observations are kept in a more active state in memory than explained observations (Baumann, 2001). However, a more comprehensive account of diagnostic reasoning will also have to incorporate predictions about the representation of already retrieved explanations and their influence on memory activation over time.

A key aspect of such considerations has to be the contrast between limited human working memory capacity and the large number of observations and explanations that might have to be maintained during diagnostic reasoning tasks. In our experiments participants had to maintain up to four symptoms in working memory; an amount that lies within the accepted range of 4 ± 1 (Cowan, 2001). However assuming that participants also store retrieved candidate explanations in memory, one would quickly reach capacity limits. Furthermore, in most real life diagnostic reasoning tasks, a higher amount of observations needs to be explained. An interesting question for further research will be to investigate what happens if the amount of information to be actively maintained during the task exceeds working memory capacity. In such a case, the least activated information might be dropped from working memory (Chuderski, Stettner, & Orzechowski, 2006; Thomas, et al., 2008) and therefore

should lose its ability to spread activation to long-term memory, unless they are actively recovered from long-term memory.

Also time and task constraints will be more complex in many real world settings. In our experiments, symptoms were presented at a fixed rate, with a relatively small spacing over time, and with (almost) no interference from other tasks. It has been proposed that information will be held by a cognitive resource like working memory until the resource is needed for another task (Salvucci & Taatgen, 2008). Applied to diagnostic reasoning as proposed in this paper, this would mean that observed symptoms would remain in working memory, until working memory is needed for something else (see also Berman, et al., 2009). With increasing spacing of the symptoms over time, and with increasing complexity of the diagnostic situation, the chance for interfering working memory use grows. Consequently, the probability for observed symptoms to be lost from working memory also grows under these conditions. Also in this case, symptoms would have to be actively recovered from long-term memory before they could affect memory activation again.

Another open question is related to the representation of knowledge in long-term memory. As we discussed in the introduction, memory activation processes can only then provide the reasoner with an adaptive set of possible explanations if diagnostic knowledge is represented in a way that fits the requirements of the task. Memory activation might for example favor the retrieval of an explanation that has been successfully used in the past compared to the retrieval of an explanation that has rarely occurred in the reasoner's experience but fits the current patient better. The representation of knowledge in long-term memory will most probably vary depending on the task structure and the way in which it was learned. In our experiments, the task structure was clearly defined and the knowledge was learned in an explicit semantic fashion through a series of practice trials. This simplification of knowledge acquisition compared to real life situations allowed us to focus on the effects of memory activation by keeping the effects of knowledge representation relatively constant. It

will be an interesting question for future research to investigate the role of different ways of knowledge representation on memory activation processes. By proposing an episodic as well as a semantic representation and specifying the memory activation processes related to these representations (Thomas, et al., 2008) already made an important step into this direction. We suspect, however, that a more detailed investigating of different ways of knowledge representation will not question the implications of our findings. A less clearly defined task structure and a more implicit acquisition of knowledge as they would be expected to occur in real life will only increase the importance of memory activation processes (Dijksterhuis & Nordgren, 2006).

Conclusion

To conclude, our results support the assumption that automatic memory activation can adaptively regulate the availability of explanations in memory and thereby provide the reasoner with a subset of explanations that have a high probability of being relevant in the current context. This regulation of explanations' availability was not only evident at the point of the diagnosis, but throughout the whole reasoning process. Future research must show whether simple models of memory activation as we tested them in this paper, prove to be sufficient to explain memory processes in real world diagnostic reasoning tasks. Further research is also needed to investigate how such simple memory models can be extended into more comprehensive models of diagnostic reasoning that take into account the interaction and respective contribution of automatic memory activation and deliberate reasoning strategies.

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.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language, 38*, 341-380.

Anderson, J. R., & Lebiere, C. (1998). The atomic components of thought. Mahway: NJ: Erlbaum.

Anderson, J. R., & Schooler, L. (1991). Reflections of the environment in memory. *Psychological Science, 2*(6).

Arocha, J. F., & Patel, V. L. (1995). Construction-integration theory and clinical reasoning. In C. Weaver, S. Mannes & C. Fletcher (Eds.), *Discourse Comprehension: Essays in Honor of Walter Kintsch* (pp. 359-382): Lawrence Erlbaum Associates.

Arocha, J. F., Wang, D., & Patel, V. L. (2005). Identifying reasoning strategies in medical decision making: A methodological guide. *Journal of Biomedical Informatics, 38*, 154–171.

Barrows, H., Norman, G., Neufeld, V., & Feightner, J. (1982). The clinical reasoning of randomly selected physicians in general medical practice. *Clinical and investigative medicine, 5*(1), 49-55.

Baumann, M. R. K. (2001). *Die Funktion des Arbeitsgedächtnisses beim abduktiven Schließen: Experimente zur Verfügbarkeit der mentalen Repräsentation erklärter und nicht erklärter Beobachtungen.* Doctoral dissertation. Chemnitz University of Technology, Chemnitz. Retrieved January 17, 2010, from http://archiv.tu-chemnitz.de/pub/2001/0071.

Baumann, M. R. K., Krems, J. F., & Ritter, F. E. (2010). Learning from examples does not prevent order effects in belief revision. *Thinking & Reasoning, 16*(2), 98-130.

Berman, M., Jonides, J., & Lewis, R. (2009). In search of decay in verbal short-term memory. *Learning, Memory, 35*(2), 317-333.

Böhm, U., & Mehlhorn, K. (2009). The Influence of Spreading Activation on Memory Retrieval in Sequential Diagnostic Reasoning. In A. Howes, D. Peebles & R. Cooper (Eds.), *Proceedings of the 9th International Conference on Cognitive Modeling*. Manchester, UK.

Borst, J., Taatgen, N., & van Rijn, H. (2010). The Problem State: A Cognitive Bottleneck in Multitasking. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36*(2), 363-382.

Bylander, T., Allemang, D., Tanner, M., & Josephson, J. (1991). The computational complexity of abduction. *Artificial Intelligence, 49*(1-3), 25-60.

Chechile, R. A. (2003). Mathematical Tools for Hazard Function Analysis. *Journal of Mathematical Psychology, 47*(5-6), 478-494.

Chinn, C., & Brewer, W. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching, 35*(6), 623-654.

Chuderski, A., Stettner, Z., & Orzechowski, J. (2006). Modeling individual differences in a working memory search task. *Proceedings of the 7th International Conference on Cognitive Modelling*. Trieste, Italy.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(01), 87-114.

Dijksterhuis, A., & Nordgren, L. (2006). A theory of unconscious thought. *Perspectives on Psychological Science, 1*(2), 95.

Dougherty, M. R. P., Gettys, C. F., & Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes, 70*, 135-148.

Dougherty, M. R. P., & Hunter, J. E. (2003a). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica, 113*(3), 263-282.

Dougherty, M. R. P., & Hunter, J. E. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition, 31*(6), 968-982.

Dougherty, M. R. P., & Sprenger, A. (2006). The influence of improper sets of information on judgment: How irrelevant information can bias judged probability. *Journal of Experimental Psychology: General, 135*(2), 262.

Dougherty, M. R. P., Thomas, R., & Lange, N. (2010). Toward an Integrative Theory of Hypothesis Generation, Probability Judgment, and Hypothesis Testing *Psychology of Learning and Motivation* (Vol. Volume 52, pp. 299-342): Academic Press.

Drewitz, U., & Thüring, M. (2009). Modeling the confidence of predictions: A time based approach. In A. Howes, D. Peebles & R. Cooper (Eds.), *In Proceedings of the 9th International Conference of Cognitive Modeling*. Manchester, UK.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Ericsson, K., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*(2), 211-244.

Evans, J. S. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review, 13*(3), 378–395.

Gettys, C., Pliske, R., Manning, C., & Casey, J. (1987). An evaluation of human act generation performance. *Organizational Behavior and Human Decision Processes, 39*(1), 23-51.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*(4), 506-528.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*(1), 1.

Johnson, T. R., & Krems, J. F. (2001). Use of current explanations in multicausal abductive reasoning. *Cognitive Science, 25*(6), 903.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The Mind and Brain of Short-Term Memory. *Annual Review of Psychology, 59*(1), 193-224.

Joseph, G., & Patel, V. (1990). Domain Knowledge and Hypothesis Genenation in Diagnostic Reasoning. *Medical Decision Making, 10*(1), 31.

Josephson, J. R., & Josephson, S. G. (1996). *Abductive inference: Computation, philosophy, technology*. New York, NY, US: Cambridge University Press.

Just, M., & Carpenter, P. (1987). *The psychology of reading and language comprehension*: Allyn and Bacon Boston, MA.

Kim, N., & Keil, F. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory and Cognition, 31*(1), 155-165.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*: Cambridge University Press.

Lewandowsky, S., Oberauer, K., & Brown, G. D. A. (2009). No temporal decay in verbal short-term memory. *Trends in cognitive sciences, 13*(3), 120-126.

Lovett, M. C., Daily, L. Z., & Reder, L. M. (2000). A source activation theory of working memory: cross-task prediction of performance in ACT-R. *Cognitive Systems Research, 1*, 99-118.

Marewski, J. N., & Olsson, H. (2009). Beyond the null ritual: Formal modeling of psychological processes. *Journal of Psychology, 217*, 49–60.

Mehle, T. (1982). Hypothesis generation in an automobile malfunction inference task. *Acta Psychologica, 52*(1-2), 87-106.

Mehlhorn, K., & Jahn, G. (2009). Modeling sequential information integration with parallel constraint satisfaction. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2469-2474). Austin, TX: Cognitive Science Society.

Meyer, D., & Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*(2), 227-234.

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in immediate serial recall: decay, temporal distinctiveness, or interference? *Psychological Review, 115*(3), 544-576.

Pitt, M., Myung, I., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*(3), 472-491.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*(2), 358-367.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237.

Salvucci, D., & Taatgen, N. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review, 115*(1), 101-130.

Sprenger, A. (2007). *Sequential hypothesis generation.* Doctoral dissertation, University of Maryland, College Park. Retrieved March 20, 2010, from ProQuest Digital Dissertations database (Publication No. AAT 3260299).

Sprenger, A., & Dougherty, M. R. (2006). Differences between probability and frequency judgments: The role of individual differences in working memory capacity. *Organizational Behavior and Human Decision Processes, 99*(2), 202-211.

Taatgen, N., & Anderson, J. (2002). Why do children learn to say 'broke'? A model of learning the past tense without feedback. *Cognition, 86*(2), 123-155.

Thagard, P. (1989a). Explanatory coherence. *Behavioral and Brain Sciences, 12*(03), 435-467.

Thagard, P. (1989b). Extending explanatory coherence. *Behavioral and Brain Sciences, 12*(03), 490-502.

Thagard, P. (2000). Probabilistic networks and explanatory coherence. *Cognitive Science Quarterly, 1*(1), 91-114.

Thomas, R. P., Dougherty, M. R. P., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review, 115*(1), 155-185.

Vallesi, A., Shallice, T., & Walsh, V. (2007). Role of the prefrontal cortex in the foreperiod effect: TMS evidence for dual mechanisms in temporal preparation. *Cerebral Cortex, 17*(2), 466-474.

Wang, H., Johnson, T., & Zhang, J. (2006a). A hybrid system of abductive tactical decision making. *International Journal of Hybrid Intelligent Systems, 3*(1), 23-33.

Wang, H., Johnson, T., & Zhang, J. (2006b). The order effect in human abductive reasoning: an empirical and computational study. *Journal of Experimental & Theoretical Artificial Intelligence, 18*(2), 215-247.

Weber, E., Böckenholt, U., Hilton, D., & Wallace, B. (1993). Determinants of diagnostic hypothesis generation: Effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(5), 1151-1164.

Author Note

Katja Mehlhorn, Department of Psychology, University of Groningen, the Netherlands; Niels A. Taatgen, Department of Artificial Intelligence, University of Groningen, the Netherlands; Christian Lebiere, Department of Psychology, Carnegie Mellon University, USA; Josef Krems, Department of Psychology, Chemnitz University of Technology, Germany.

This research was supported by a research grant from the German Academic Exchange Service (DAAD) awarded to Katja Mehlhorn.

Correspondence concerning this article should be addressed to Katja Mehlhorn, Department of Psychology, University of Groningen, Postbus 72, 9700 AB Groningen, the Netherlands. Email: S.K.Mehlhorn@rug.nl.

Footnotes

[1] In contrast to the concept of spreading activation between positively associated memory elements, the concept of spreading inhibition between negatively associated memory elements is neglected in many theories of memory retrieval as it often has little practical impact (cf. Anderson & Lebiere, 1998). However, in diagnostic reasoning making a certain observation does not only increase the probability for positively associated explanations to be the correct diagnosis, but it also decreases the probability of other explanations. Consequently, inhibition between observations and non-supported explanations becomes important (Dougherty & Sprenger, 2006).

[2] While manipulating uncertainty in this way represents a strong simplification of real life diagnostic uncertainty, we chose for this design for two main reasons. First, varying the position of the unreliable information within trials would have required a far larger number of trials. The number of trials already being very large, we decided against this (potentially very interesting) manipulation. Second, not informing participants about the potential unreliability of the second symptom might have resulted in a variety of potential strategies in dealing with incoherent trials (see Chinn & Brewer, 1998 for an overview of potential strategies in dealing with incoherent data). By informing participants which symptom might be unreliable, we attempted to reduce the amount of possible strategies.

[3] In incoherent trials a third type of target probe was used (rejected targets). Rejected targets probed explanations that were compatible with the first but incompatible with the second symptoms. The reactions to those probes were in line with our predictions. However, as those probes were only presented in the incoherent trials, we will not report them here.

[4] To test for the robustness of our findings, we also conducted all analyses of the reaction time data based on the medians (without excluding outlier values). The primary results are consistent across analyses.

[5] All ANOVAS were repeated-measures ANOVAS.

[6] Bayes Factors larger than 1.0 are taken as evidence in factor of the null whereas BF < 1.0 are taken as evidence in favor of the alternative. See Rouder, Speckman, Sun, Morey, & Iverson (2009) for derivations and a guide for interpreting the magnitude of Bayes Factors.

[7] Whereas the result of higher diagnosis times in coherent trials might seem counterintuitive, it is most likely caused by the number of symptoms presented before the diagnosis, rather than by the coherence of the trial. Incoherent trials always consisted of four symptoms, whereas coherent trials could consist of three (56% of all coherent trials) or four (44% of all coherent trials) symptoms. Analyzing coherent three-symptom and four-symptom trials separately shows that coherent four-symptom trials were in general responded to faster than coherent three-symptom trials ($mean_{four}$ = 644 ms (*SD* = 208); $mean_{three}$ = 915 ms (*SD* = 223); $t(21) = 11.233$, $p < .001$) and that the diagnosis times in coherent four-symptom trials were significantly faster than in incoherent trials, $t(21) = 4.684$, $p < .001$.

[8] To further test if participants indeed ignored the second symptom, we compared the diagnostic performance in coherent trials where ignoring the second symptom allowed for unambiguously finding the correct diagnosis (unambiguous coherent trials) and coherent trials where ignoring the second symptom did not allow for finding the correct diagnosis (ambiguous coherent trials). Indeed diagnosis accuracy was marginally higher in unambiguous (*mean* = 95.8%; *SD* = 3.8) than in ambiguous coherent trials (*mean* = 93.8%; *SD* = 7.4), $t(21)=1.815$, $p = .084$. Diagnosis times for correct diagnoses were considerably faster in unambiguous (*mean* = 757 ms; *SD* = 195) than in ambiguous coherent trials (*mean* = 1053 ms; *SD* = 363), $t(21)=5.297$, $p < .001$, suggesting that participants used time at the end of the trial to solve the ambiguity caused by ignoring the second symptom.

[9] Building the ignoring of misleading information and the increasing response preparedness into the models allowed us for assessing whether the response pattern indeed could have been caused by the interaction of memory activation and these task specific factors. It is important to note however, that these additional model components alone would not have been possible

to fit the participants' responses. Without an effect of observations on reaction times to the probes, ignoring the second symptom would not predict any effect on the reaction time data by itself. Increasing response preparedness alone would predict a decrease of reaction times over the trial, but no differences or interactions between the different probe types.

[10] To model working memory we use one of the buffers of ACT-R's cognitive modules, the imaginal buffer. The imaginal buffer is commonly used to hold a mental representation of the problem currently in the focus of attention (Borst, Taatgen, & van Rijn, 2010).

[11] For being able to directly compare the levels of explanations' availability over the course of the trial, we kept the total amount of the scaling parameter W constant after the fourth symptom of the trial. This choice was somewhat arbitrary, as we could have kept W constant at any other point during the trial (e.g., using a constant value W_1 after the first symptom of the trials). Note however, that this would have not changed the results substantially, as it would have merely produced a linear transformation of all scaling values. To test this we implementing all models with a constant value of W_1=.16. This produced the same pattern over the course of the trial, however, with much smaller differences between the different probe types at each point during the trial; leading to much smaller values for R^2 and lower diagnosis accuracies for all models, but *Model-Number*.

[12] Note that not only the chemicals but also the foils are represented in memory. This is because, contrary to lexical decision tasks, where a constrained number of words stands against an unconstrained number of non-words, in our experiment chemicals and foils each consisted of a set of nine letters which were taught to the participants in the training session.

[13] Reflecting the probabilities for upcoming stimuli, the base-level activations of the expectations vary. As probes are presented equally often after one of the four symptoms, the probability of a probe to be presented after the first symptom is only .25. Consequently, the base-level of an expect-probe chunk after the first symptom is so much lower than the base-level of an expect-symptom chunk that the model will retrieve an expect-probe chunk only in

about 25% of all trials. With each additional symptom that is presented without a probe, the probability of a probe (reflected by the base-levels of the expect-probe chunks) increases (to .33, .5, and 1 respectively). Consequently, the earlier in the trial the probe appears, the higher the chance that the model retrieves no expect-probe chunk and has to make a time-costly change to its expectation. The model changes its expectation by firing an additional production rule (costing 50 ms).

[14] ACT-R's latency factor (F) was set to 1.4 and activation noise (s) to .05. All facts in memory were set to equal, relatively high base-levels of 2, modeling trained participants. Positive associative strengths (S_{ji}) were calculated using Equation 6, with the maximum associative strength (S) set to 2.5. Negative associative strengths (S_{ji}) were estimated from the data to be -.75. The total amount of W that the models spread after four symptoms were presented was set to .48.

[15] As no symptoms were ignored the models now had one more symptom to integrate than in Experiment 1. Therefore, the total amount of W that the models spread after the fourth symptom was set to .64.

Table 1

Summary of the Material Participants had to Learn in Experiment 1.

Aggregate state and source of contamination	Category	Chemical	Specific symptoms		Unspecific symptoms		
gasiform --- inhaled	Landin	B	cough	short breath	headache	eye inflammation	
		T	cough	vomiting	headache		itching
		W	cough			eye inflammation	itching
crystalline --- skin contact	Amid	Q	skin irritation	redness	headache	eye inflammation	
		M	skin irritation	short breath	headache		itching
		G	skin irritation			eye inflammation	itching
liquid --- drinking water	Fenton	K	diarrhea	vomiting	headache	eye inflammation	
		H	diarrhea	redness	headache		itching
		P	diarrhea			eye inflammation	itching

Note. Original material in German.

Table 2

Coherent and Incoherent Sample Trial for Experiment 1.

Order	Symptoms	Explanations supported by current symptom	Possible target probes		Possible foils
			Compatible	Incompatible	
Coherent trial					
1st	cough	BTW	BTW	QMGKHP	FZVDNCXLR
2nd	vomiting	TK	T	QMGKHP	FZVDNCXLR
3rd	itching	TWMGHP	T	QMGKHP	FZVDNCXLR
4th	headache	BTQMKH	T	QMGKHP	FZVDNCXLR
Correct diagnosis: T					
Incoherent trial					
1st	cough	BTW	BTW	QMGKHP	FZVDNCXLR
2nd	red eyes	WG	W	QMGKHP	FZVDNCXLR
3rd	short breath	BM	B	QMGKHP	FZVDNCXLR
4th	headache	BTQMKH	B	QMGKHP	FZVDNCXLR
Correct diagnosis: B					

Note. Shown for each symptom: Supported explanations, possible target probes, and foils. Note that the set of potential incompatible probes stays the same over the trial (it consists of those explanations that are not supported by the first symptom) while the set of potential compatible probes changes as the number of explanations supported by all symptoms decreases.

Table 3

Results of the ANOVAs for compatible targets, incompatible targets and foils after each symptom in Experiment 1.

Effect	Factors	*F*	*p*	η_p^2
Interaction	Type of Probe (compatible, incompatible, foil) x Symptoms before Probe (one, two, three, four)	(2,42) = 5.03	**.011**	.19
Interaction	Type of Probe (compatible, incompatible) x Symptoms before Probe (one, two, three, four)	(1,21) = 7.15	**.014**	.25
Interaction	Type of Probe (compatible, foil) x Symptoms before Probe (one, two, three, four)	(1,21) = 2.35	.140	.10
Main effect	Type of Probe (compatible, foil)	(1,21) = 4.49	**.046**	.18
Interaction	Type of Probe (incompatible, foil) x Symptoms before Probe (one, two, three, four)	(1,21) = 3.70	**.068**	.15
Simple effect for compatible	Symptoms before Probe (one, two, three, four)	(1,12) = 20.21	**< .001**	.49
Simple effect for incompatible	Symptoms before Probe (one, two, three, four)	(1,21) = 0.46	**.506**	.02
Simple effect for foil	Symptoms before Probe (one, two, three, four)	(1,21) = 25.56	**< .001**	.55
Simple effect after symptom 1	Type of Probe (compatible, incompatible, foil)	(2,42) = 12.49	**< .001**	.37
Simple effect after symptom 2	Type of Probe (compatible, incompatible, foil)	(2,42) = 1.21	.309	.05
Simple effect after symptom 3	Type of Probe (compatible, incompatible, foil)	(2,42) = 29.13	**< .001**	.58
Simple effect after symptom 4	Type of Probe (compatible, incompatible, foil)	(2,42) = 17.41	**< .001**	.45

Note. p-values < .1 are indicated in bold. For non-significant interactions the main effect of Type of Probe is reported additionally.

Table 4

Fits for probe reaction times (R^2 and RMSD) and the diagnostic performance (accuracy and reaction time) of each model for Experiment 1 and Experiment 2.

	R^2	RMSD (ms)	Diagnosis Accuracy (%)	Diagnosis Time (ms)
Experiment 1				
Human Data (SD)			95.5 (3.7)	705 (167)
Model-Current	.79	30	28	586
Model-Time	.79	28	53	597
Model-Constant	.70	38	86	592
Model-Number	**.85**	**27**	**85**	**569**
Experiment 2				
Human Data (SD)			95.9 (3.9)	574 (264)
Model-Current	.24	61	27	566
Model-Time	.37	75	71	584
Model-Constant	.45	60	95	587
Model-Number	**.71**	**83**	**92**	**589**

Note. The best fitting model is indicated in bold.

Table 5

Summary of the Material Participants had to Learn in Experiment 2.

Aggregate state and source of contamination	Category	Chemical	Specific symptoms		Unspecific symptoms		
gasiform --- inhaled	Landin	B	cough	short breath	headache	eye inflammation	
		T	cough	short breath	headache		itching
		W	cough			eye inflammation	itching
crystalline --- skin contact	Amid	Q	skin irritation	redness	headache	eye inflammation	
		M	skin irritation	redness	headache		itching
		G	skin irritation			eye inflammation	itching
liquid --- drinking water	Fenton	K	diarrhea	vomiting	headache	eye inflammation	
		H	diarrhea	vomiting	headache		itching
		P	diarrhea			eye inflammation	itching

Note. Original material in German.

Table 6

Sample Trial for Experiment 2.

Order	Symptoms	Explanations supported by current symptom	Possible target probes				
			Compatible	In-compatible	Rejected-after-2	Rejected-after-3	Rejected-after-4
1st	headache	BTQMKH	BTQMKH	WGP	---	---	---
2nd	cough	BTW	BT	WGP	QMKH	---	---
3rd	short breath	BT	BT	WGP	QMKH	-	---
4th	itching	TWMGHP	T	WGP	QMKH	-	B
Correct diagnosis: T							

Note. Shown for each symptom: Supported explanations and possible target probes ("---" marks cells that cannot be filled in general; "-" marks cells that cannot be filled in this particular trial). Foils were identical to Experiment 1.

Table 7

Results of the ANOVAs for compatible targets, incompatible targets and foils after each symptom in Experiment 2.

Effect	Factors	*F*	*p*	η_p^2
Interaction	Type of Probe (compatible, incompatible, foil) x Symptoms before Probe (one, two, three, four)	(2,50) = 3.84	**.028**	.13
Interaction	Type of Probe (compatible, incompatible) x Symptoms before Probe (one, two, three, four)	(1,25) = 5.65	**.025**	.19
Interaction	Type of Probe (compatible, foil) x Symptoms before Probe (one, two, three, four)	(1,25) = 0.90	.352	.04
Main effect	Type of Probe (compatible, foil)	(1,25) = 10.88	**.003**	.30
Interaction	Type of Probe (incompatible, foil) x Symptoms before Probe (one, two, three, four)	(1,25) = 3.39	**.077**	.12
Simple effect for compatible	Symptoms before Probe (one, two, three, four)	(1,25) = 34.46	**< .001**	.58
Simple effect for incompatible	Symptoms before Probe (one, two, three, four)	(1,25) = 9.49	**.005**	.28
Simple effect for foil	Symptoms before Probe (one, two, three, four)	(1,25) = 68.46	**< .001**	.73
Simple effect after symptom 1	Type of Probe (compatible, incompatible, foil)	(2,50) = 4.37	**.018**	.15
Simple effect after symptom 2	Type of Probe (compatible, incompatible, foil)	(2,50) = 2.10	.133	.08
Simple effect after symptom 3	Type of Probe (compatible, incompatible, foil)	(2,50) = 3.76	**.030**	.13
Simple effect after symptom 4	Type of Probe (compatible, incompatible, foil)	(2,50) = 7.60	**.001**	.23

Note. p-values < .1 are indicated in bold. For non-significant interactions the main effect of Type of Probe is reported additionally.

Table 8

Results of the ANOVAs for rejected-after-2 targets, compatible targets, incompatible targets, and foils after symptom two, three, and, four in Experiment 2.

Effect	Factors	*F*	*p*	η_p^2
Interaction	Type of Probe (rejected-after-2, compatible, incompatible, foil) x Symptoms before Probe (two, three, four)	(3,75) = 1.89	.138	.07
Main effect	Type of Probe (rejected-after-2, compatible, incompatible, foil)	(3,75) = 8.44	**< .001**	.25
Interaction	Type of Probe (rejected-after-2, compatible) x Symptoms before Probe (two, three, four)	(1,25) = 4.52	**.043**	.15
Interaction	Type of Probe (rejected-after-2, Incompatible) x Symptoms before Probe (two, three, four)	(1,25) < .01	.980	< .01
Main effect	Type of Probe (rejected-after-2, Incompatible)	(1,25) =.06	.811	< .01
Interaction	Type of Probe (rejected-after-2, foil) x Symptoms before Probe (two, three, four)	(1,25) = 1.20	.284	.05
Main effect	Type of Probe (rejected-after-2, foil)	(1,25) =.06	.149	.08
Simple effect for rejected-after-2	Symptoms before Probe (two, three, four)	(1,25) = 5.80	**.024**	.19
Simple effect after symptom 2	Type of Probe (rejected-after-2, compatible, incompatible, foil)	(3,75) = 2.09	.108	.08
Simple effect after symptom 3	Type of Probe (rejected-after-2, compatible, incompatible, foil)	(3,75) = 2.70	**.052**	.10
Simple effect after symptom 4	Type of Probe (rejected-after-2, compatible, incompatible, foil)	(3,75) = 6.91	**< .001**	.22

Note. p-values < .1 are indicated in bold. For non-significant interactions the main effect of Type of Probe is reported additionally.

Figure Captions

Figure 1. Illustration of the trial-procedure for a sample trial from Experiment 1.

Figure 2. Mean (±1 SE) reaction time to probes over the course of trials in Experiment 1. Human data and model data. The models will be described later in the text.

Figure 3. Mean (±1 SE) reaction time to probes over the course of trials in Experiment 2. Human data and model predictions.

Figure 1.

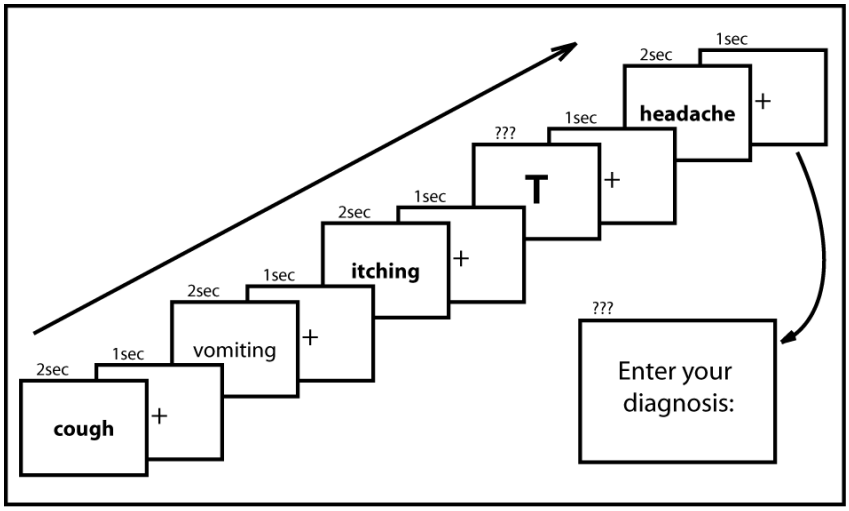
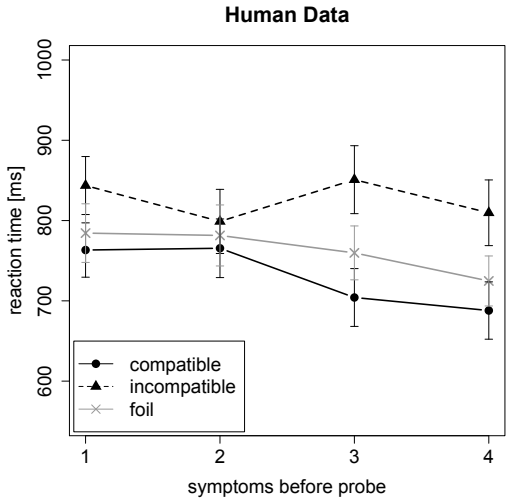

2sec
1sec
cough
+
2sec
1sec
vomiting
+
2sec
1sec
itching
+
???
1sec
T
+
2sec
1sec
headache
+
???
Enter your
diagnosis:

Figure 2.

2a) Human Data



2b) Model Data

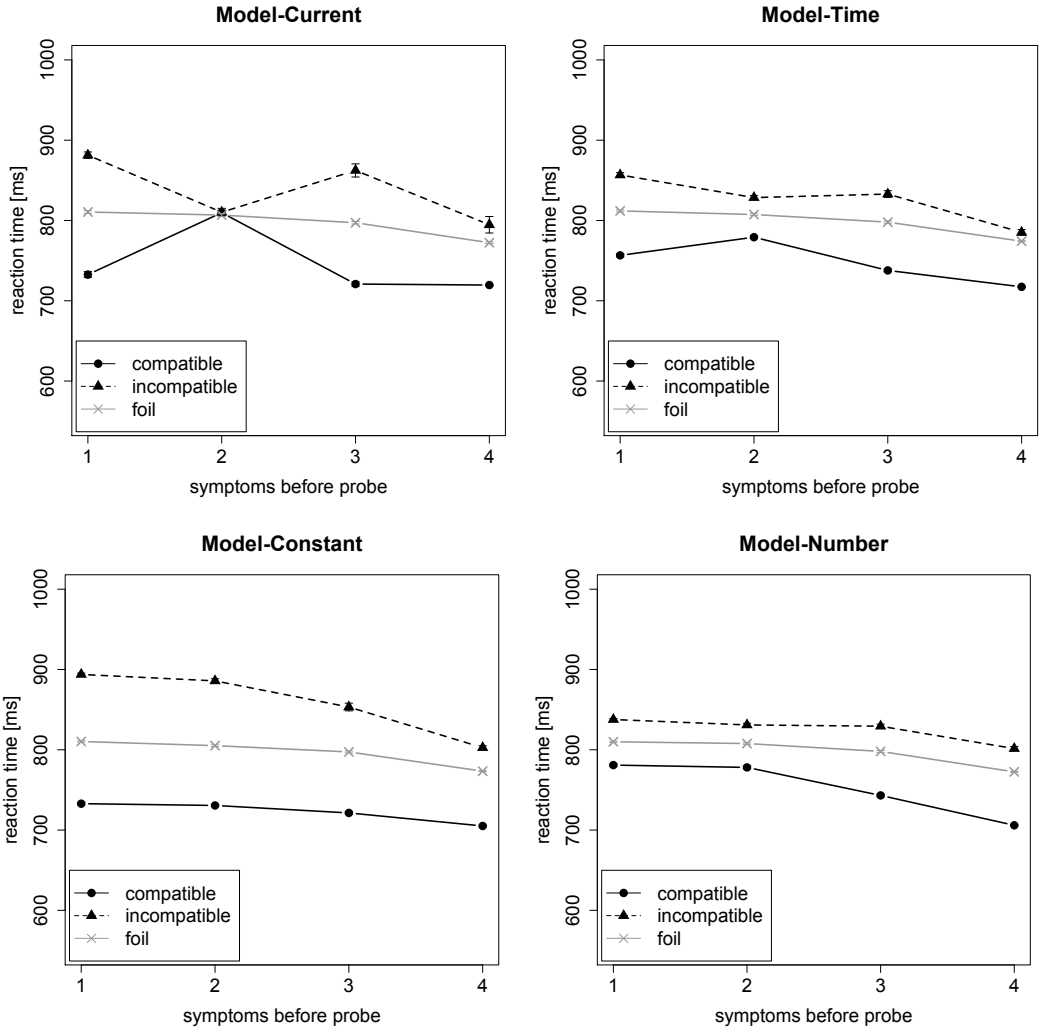
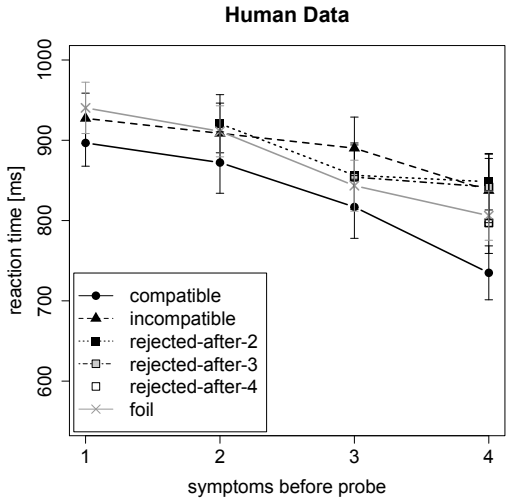


Figure 3.

3a) Human Data



3b) Model Predictions