Cognitive modeling of anticipation: Unsupervised learning and symbolic modeling of pilots' mental representations

Sebastian Blum[1*], Oliver Klaproth[2†] and Nele Russwinkel[1‡]

[1]*Department of Cognitive Modeling in Dynamic Human-Machine Systems, TU Berlin, Marchstr. 23, 10587 Berlin, Germany*
[2]*Central Research & Technology, Airbus, Hein-Sass-Weg 22, 21129 Hamburg, Germany*

Abstract

The ability to anticipate team members' actions enables joint action towards a common goal. Task knowledge and mental simulation allow for anticipating other agents' actions and for making inferences about their underlying mental representations. In human-AI teams, providing AI agents with anticipatory mechanisms can facilitate collaboration and successful execution of joint action. This paper presents a computational cognitive model demonstrating mental simulation of operators' mental models of a situation and anticipation of their behavior. The work proposes two successive steps: (1) A hierarchical cluster algorithm is applied to recognize patterns of behavior among pilots. These behavioral clusters are used to derive commonalities in situation models from empirical data (N = 13 pilots). (2) An ACT-R cognitive model is implemented to mentally simulate different possible outcomes of action decisions and timing of a pilot. Model-tracing of ACT-R allows following up on operators' individual actions. Two models are implemented using the symbolic representations of ACT-R: One simulating normative behavior and the other by

*E-mail: sebast.blum@gmail.com
†E-mail: Oliver.Klaproth@airbus.com
‡E-mail: Nele.Russwinkel@tu-berlin.de

simulating individual differences and using subsymbolic learning. Model performance is analyzed by a comparison of both models. Results indicate improved performance of the individual differences over the normative model and are discussed regarding implications for cognitive assistance capable of anticipating operator behavior.

Key words: cognitive modeling, model-based cognitive assistance, anticipation, mental simulation, model-tracing, unsupervised machine learning

1 Introduction

In complex and dynamic situations, operators of machines and computers benefit from assistance of artificial intelligence (AI). Virtual assistants in commercially available products are often sensitive to the changing characteristics of a task or situation, for example changing speed limits or light conditions in the automotive domain. However, operators' cognitive states aside from fatigue (Schmidt, Braunagel, Stolzmann, & Karrer-Gauss, 2016) are largely ignored. Tailoring assistance to operator's cognitive processes is an active research field that attracts interest from various scientific areas and industries (Flumeri et al., 2019; Scharfe & Russwinkel, 2019; Teo, Matthews, Reinerman-Jones, & Barber, 2020). Cognitive assistance that supports the operator and reduces demands is likely to bring performance improvements in human-machine interaction, especially in safety-critical domains with high cognitive demands and complex automation such as aviation (Estes et al., 2016).

1.1 Cognitive assistance in aviation

While most automation is designed to assist and to decrease workload, it poses additional demands on operators. Operators need to adapt to such systems which causes additional workload, for example by the necessity to be monitored (Bainbridge, 1983). They need to pay attention to and integrate information from various sources to build up an understanding of the situation and to be prepared in cases of failure or otherwise unexpected events. This information acquisition and integration, also referred to as situation assessment and awareness (Endsley, 1995), is established in three stages: (1) perception, (2) comprehension, and (3) projection of events. As an example, pilots develop situation awareness in emergency situations by perceiving (I) an alert sound and warning message, by comprehending (II) their meaning for the situation, for example the malfunction of a system and by projecting (III) near future developments and necessary preparations. This situation awareness, also referred to as a situation model, represents a basis for operators' decision making (Endsley, 1995).

In commercial aviation, the pilot monitoring and pilot flying can act as living cognitive assistants to one another (Estes et al., 2016), for example by inferring the other's understanding of a situation. In line with Estes et al. (2016), we expect that cognitive assistance by an AI agent can support pilots in reduced crew operations. To support adequately, cognitive assistance needs to monitor pilots' perception and know what information they are aware of and what not (Klaproth, Halbrügge, et al., 2020). Based on knowledge of the task and pilot, inferences can be made about what experience and knowledge are activated by perception to form a comprehensive situation model. Emulating such empathic capabilities as inferring operators' comprehension of a situation remains a challenge for AI agents.

1.2 Mental simulation and anticipation

Comprehension of a situation is moderated by pre-existing knowledge that helps make sense of environmental stimuli, so called mental representations or models (Endsley, 1995), that can be seen as a prototypical state (Salas, 2017) or a memory representation (Furlough & Gillan, 2018) of a system or situation. Mental models are commonly made explicit through the use of qualitative methods such as interviews, but they can also be extracted from behavioral data by more quantitative techniques (e.g., Carley and Palmquist, 1992; Steiger and Steiger, 2009). Similarities in behavior can point to commonalities in mental models between operators (Hegarty et al., 1988). Situation models can be understood as an instantiation of mental models, which represent operators' understanding of the present state of the system (Endsley, 2000). In collaborative settings, shared mental or situation models, anticipation, and the ability to prepare action based on anticipated effects are essential for effective teamwork (Sebanz, Bekkering, & Knoblich, 2006). Dynamic simulation of situation models (also referred to as mental simulation; Klein, Phillips, Rall, & Peluso, 2007) can help agents to make sense of and anticipate operators' actions and to envision their own actions in response.

Envisioning an action means to foresee and to anticipate it, and to be able to act in advance. Anticipatory systems, biological (human) or artificial, are capable of generating predictions about possible outcomes of actions by conceiving possible future states. This enables future-oriented action and behavior that is based on implicit or explicit representations of future events (Poli, 2010; Riegler, 2003). Anticipatory mechanisms can enhance behavioral capabilities of artificial systems in various areas, such as action selection, reasoning, imitation, and perspective-taking, yet even in complex forms of interaction such as trade and language (Butz & Pezzulo, 2008; Pezzulo, Butz, & Castelfranchi, 2008). The ability to anticipate operators' actions would not only support the autonomy and contextual reasoning of AI agents (Amos-Binks & Dannenhauer, 2019), but also facilitate more natural interaction with human co-operators (Pandey, Ali, & Alami, 2013).

According to Sebanz et al. (2006), this would require the AI agent (1) to be sensitive to not only the state of the aircraft and the dynamic environment, but also regarding pilot behavior. Furthermore, (2) to hold a vast amount of knowledge about how the task should ideally and will likely be performed, and (3) to be able to compare competing situation models to anticipate pilot behavior in the near future. Adapting to human actions to the end of providing cognitive assistance requires AI agents to be able to take an operator's perspective and distinguish changes caused by themselves from changes caused by others, so they can assist operators individually and before critical events arise (Butz & Pezzulo, 2008; Russwinkel, Vernaleken, & Klaproth, 2020). We assume that a cognitively-plausible computational model of anticipation should be able to compare multiple (competing) situation models when taking someone else's perspective. Cognitive architectures allow for a cognitively plausible implementation of anticipatory thinking capabilities in computational models that are grounded in empirical evidence of human cognitive processing (Amos-Binks & Dannenhauer, 2019) and have been previously used to incorporate perspective taking (Kennedy, Bugajska, Adams, Schultz, & Trafton, 2008).

1.3 Cognitive systems

Originally, understanding and reproducing intelligent behavior in cognitively plausible computational systems was the central idea of artificial intelligence (AI) (Langley, 2011, 2012). Today's AI rather focuses on one specific component that represents intelligent behavior, for example gradual learning used in machine learning. Cognitive assistance should avoid this limitation and reconnect to AI's original goal of understanding the human mind, to exhibit flexible capabilities like a human assistant. Langley (2012) proposes the term cognitive systems to describe systems that evolve around higher levels of cognition, aim for a more general nature of intelligence, and in the end enable a system to act and think like humans. We propose a computational model for cognitive assistance as they provide guidelines for creating a computational replication of high-level cognition to solve novel and complex problems (Langley, 2006). By focusing on the underlying cognitive dynamics human errors might be anticipated during the operators' comprehension of the situation and before erroneous actions are executed. This aim is different to precisely simulating human behavior.

1.4 Cognitive architectures

Cognitive architectures are comprehensive theory-driven computational models of the mind. ACT-R (Adaptive Control of Thought - Rational) is the most widely used architecture (Kotseruba & Tsotsos, 2018). It is based on a modular decomposition of the brain, associated with distinct cortical regions, e.g. perceptual-motor modules, and memory modules. The modules are linked to the core procedural module, which stores and applies procedural knowledge. ACT-R utilizes two types of knowledge: declarative knowledge and procedural knowledge, both realized in a symbolic-connectionist structure. Declarative knowledge uses chunks to store explicit facts of the system's knowledge, whereas productions (procedural knowledge) are condition-action statements that utilize those facts to modify the current state of the system. The processing of the production system and the knowledge chunks is guided by a subsymbolic activation of their structures. The chunks of the declarative knowledge incorporate an "activation", which is a measure of the chunks relevance and activity in memory. In the procedural system, a production's "utility" determines which production is selected if more than one production applies to the current system state. Both, the activation of chunks and the utility of productions account for learning within ACT-R.

ACT-R has seen many applications in aviation (e.g., Taatgen, Huss, Dickison, & Anderson, 2008; Somers and West, 2013; Schoppek & Boehm-Davis, 2004; Byrne et al., 2004) and has also been used for research on incorporating perspective-taking within a robotic teammate (Kennedy et al., 2008). Most studies using ACT-R rely on laboratory-like simulations and do not account for real-time and real-life human behavior. Intelligent tutoring systems are an example of practical applications of cognitive models that are adaptive to human behavior. These models observe student behavioral data at runtime through a process called model-tracing to make inferences about their knowledge state, for example knowledge about commutative properties when solving math problems. Fu et al. (2006) showed how model-tracing can help identify a human operator's needs for assistance in an Anti-Air Warfare Coordinator task. This idea has been transferred to an aviation context by Klaproth, Halbrügge, et al. (2020). Through model-tracing, a cognitive agent can compare human actions to normative or desired behavior behavior models and give appropriate assistance when required (Fu et al., 2006). In the frame of the synthetic teammate project, Ball et al. (2010) have built a very extensive model that serves as a pilot for unmanned aerial vehicles and a teammate to human team members able to communicate through text chat (see Demir et al., 2015). While this is an excellent example of advancing human-AI teaming with cognitive modeling, anticipation of human teammates' information needs still remains a challenge for the synthetic teammate (McNeese, Demir, Cooke, & Myers, 2018).

1.5 A cognitive systems approach with machine learning

The implementation of an AI agent providing cognitive assistance should be inspired but not limited by principles of human cognition. Whereas abstract (i.e., symbolic) knowledge representations enable humans to learn from single events through transfer learning, machine learning algorithms have the potential to learn faster and with fewer interruptions over recurring events than humans do (Silver et al., 2016). Algorithms can perform extensively more training sessions and on multiple varying scenarios if given enough data.

We suggest cognitive assistance to be trained on big data such as flight records, resulting in more "experience" of the system. This can result in a more complete description of pilot actions across a variety of situations, and lead to more detailed representations of situation models that moderate this behavior. However, machine learning is not able to account for symbolic reasoning or planning as it focuses on statistical learning. Designing assistance according to the cognitive systems paradigm should exhibit flexible capabilities of a human assistant and aim for a more general nature of intelligence.

Approaches to combine unsupervised learning within cognitive architectures have been undertaken by (Vinokurov, Lebiere, Wyatte, Herd, & O'Reilly, 2012) who propose a hybrid framework combining ACT-R with Leabra, a cognitive architecture focused on network-based learning. In their application, ACT-R acts as a metacognitive processor and directs Leabra's visual attention to identify and classify various objects. Another approach to make use of unstructured real-world data is the combination of ACT-R with machine learning algorithms by (Thomson, Lebiere, & Bennati, 2014), who proposed a bottom-up data analysis hierarchy starting with machine learning to organize unstructured data and using cognitive architectures to refine the loosely organized data to a more elaborate semantic form.

We propose the cognitive architecture ACT-R to model mental simulation and anticipation in combination with machine learning to instantiate the respective situation models. Identifying pilot behavior patterns using machine learning takes advantage of computational performance benefits to organize unstructured data and to gain patterns from pilots' data. By designing cognitive assistance with symbolic representations of ACT-R, we intend to propose a cognitive system in line with the definition of Langley (2012). The proposed cognitive assistant is designed to observe pilots' actions and compare them to stereotypical behavior in a given situation. Based on the simulation of up to three mental models (Klein et al., 2007), the assistant seeks to identify pilots' current situation models to anticipate their future actions and to support accordingly.

1.6 Current approach

In this paper, we simulate situation models about participants in an empirical flight simulator study. By making inferences about situation models, an AI agent will be able to build a mental model of the participant itself. These mental models could enable the AI agent to anticipate human pilot behavior and information needs in similar ways to how a human pilot monitoring anticipates a human pilot flying's actions. However, direct examination of a human co-pilot's mental model building and simulation is beyond the scope of this study. In the absence of empirical data from a human co-pilot that could be used for model validation, modeling of anticipation capabilities will be evaluated with regard to accuracy of behavior predictions.

With the presented concept and implementation we intend to explore three assumptions. (1) Based on Thomson et al. (2014) we suggest that a combination of machine learning and cognitive architectures proposes a computational implementation of cognitive assistance that follows human-like ways of collaboration. In line with Hegarty et al. (1988) we assume that similarities between participants' performance point to similarities in underlying situation models that can be used for anticipation of future actions. (2) Kennedy et al. (2008) have shown that mental simulation of a cognitive model by simulating another cognitive model can result in a more effective artificial teammate. We extend the simulation process to provide the model with human-lik capabilities of simulating three submodels (Klein et al., 2007) to enhance its performance compared to a model without anticipation capabilities. (3) We hypothesize that utility learning can lead to more immediate selection of fitting mental models such that pilot behavior is anticipated correctly earlier in time, and that training a model using utility learning can result in an overall higher accuracy of anticipation.

2 Methods

This work is a combination of two successive steps: (1) An unsupervised machine learning algorithm is applied to identify different groups of pilot behavior gained from a flight simulator study, and (2) an ACT-R model is simulated that employs mental simulation of human pilot's situation models to anticipate their actions. Machine learning results will be interpreted to assume commonalities in underlying situation models within each cluster.

2.1 Participants

13 (male) pilots participated in the study. Their mean age was 53.85 (*SD* = 5.64), with an average experience of 7210.00 (*SD* = 5005.48) in-flight hours. Average flight hours over the past twelve months before participation were 228.85 (*SD* = 180.93) hours. All participants held a commercial pilot license with a multi-engine and instrument rating. The majority of pilots were test pilots and flight instructors with an Airbus A350 type rating. All participants had normal or corrected to normal eye vision and were predominantly right-handed (one ambidexter). All participants had given informed consent before participation.

2.2 Flight simulator and scenario

An empirical flight simulator study was conducted at Airbus' simulator facilities in Toulouse, France. All sessions were performed in a fixed base experimental flight simulator approximating an Airbus A350 cockpit design. Essential flight instrument properties and state changes were recorded in log data with a sampling rate of 100 Hz. As participants' interactions with cockpit systems resulted in a state change of aircraft

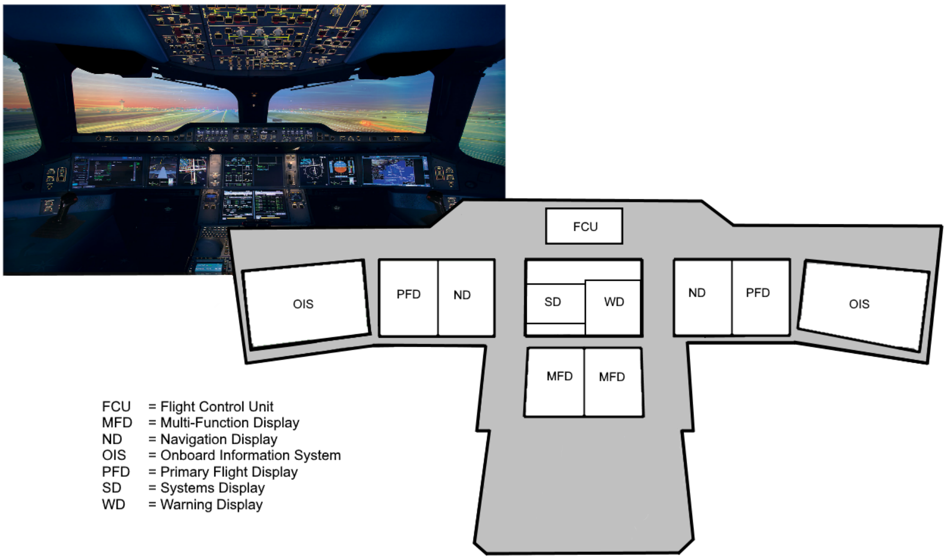


FIGURE 1: Display names for visual focus areas of the Airbus A350 cockpit layout used in the simulator study (copyright: Airbus).

systems, their inputs to controls were derived from the log data. A mobile eye-tracking headset (SMI Eye-Tracking Glasses 2) was used to record the visual focus of participants during flight with a sampling rate of 60 Hz. Focus areas were coded according to displays of the Airbus A350 cockpit design (seeFigure 1).

Participants had to perform five flight scenarios that consisted of a total of 14 events resembling real-world flight deck alerts. Each event was preceded by an auditory alert. Participants were seated in the captain's (left) seat and had to perform the task as a pilot flying. All participants were supported by the same first officer as the pilot monitoring. Events had to be resolved by a sequence of actions specified in the alert checklist or standard operating procedures, representing normative behavior in response to given alerts. From these 14 events, the engine fire event showed the most coherent sequence of recordable pilot actions, resulting in the most well separable cluster groups, and was therefore selected as primary focus in this work. In the empirical study, normative behavior in response to the engine fire consisted of hearing and processing the alert tone, looking at the warning display (WD) and reading the displayed alert message, setting the Thrust Lever of the affected engine to idle, and switching off the Engine Master. The actions were recorded in the simulator log data and analyzed to recognize behavior patterns among participants. Due to measurement issues and discomfort/visual obstruction expressed by some participants, complete datasets including eye-tracking records in the engine fire event were obtained for eight out of the total 13 pilots participating in the simulator study[1] .

2.3 Unsupervised machine learning

An agglomerative hierarchical cluster analysis (HCA) using the complete-linkage method was performed to identify patterns between pilots' behavior and select the optimal number of distinct groups, using the language R (v3.5.1) and the stats package (v3.6.2; hclust). As no differences in correctness of performing the task were observed, the cluster analysis was based on quantitative differences between the response times (RT) in the log data and on qualitative differences of the visual focus gained by eye-tracking. Pilots' visual focus is denoted by their direction of gaze before and during the event. The RT denotes the time a pilot takes to act, for example to set the Thrust Lever of the left engine to idle or to switch off the Engine Master. A regression analysis was conducted to examine potential relationships between the pilots' behavioral differences and their experience in flight hours.

2.4 Cognitive modeling

Mental simulation of a pilot's actions is done by a distinction between the cognitive model itself and the simulated dynamics of a situation model. A main cognitive model is imple-

[1]Since the modeling approach described in 2.5 was unable to cope with overlap between clusters of pilot behavior, data of one participant had to be excluded from further analysis. This resulted in clusters whose behavior was internally consistent and exclusively different from each other.

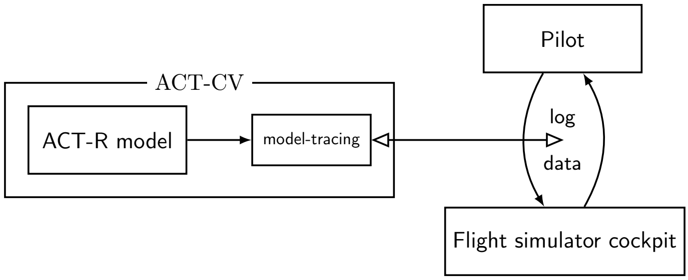


FIGURE 2: The cognitive model interacts during runtime with simulator log data containing simulator states and pilot actions in a process known as model-tracing.

mented (metamodel) which represents the cognitive assistant and simulates each separate situation model (submodel) that could apply in the given situation. Each model, metamodel or submodel, is implemented as a separate ACT-R model. Each submodel contains the visual focus and action response time of the respective situation model, implemented with ACT-R's production system. The metamodel can trigger the simulation of a submodel's production and compare the resulting actions to the observed pilot actions. Model-tracing is used to monitor a pilot's interaction with a system and a system's states. ACT-CV Halbruegge (2013) has been adapted to Python and extends ACT-R with a visual environment to establish a connection between ACT-R and flight simulator data (see Figure 2). The metamodel can thus observe a pilot's actions via chunks of the visicon. A framework denoted as ACT-MS (ACT-R Mental Simulation) has been implemented to coordinate the interaction between multiple ACT-R models[2] (see Figure 3). It is based on the Python dispatcher included in the ACT-R Version 7.13, to schedule buffer modifications within python functions and thereby to exchange parameters and variables between models.

[2] see https://pypi.org/project/pyactms/ and https://pypi.org/project/pyactcv/

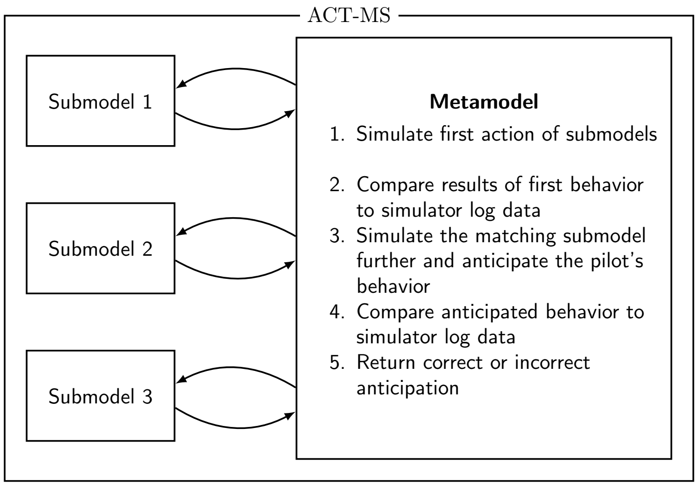


FIGURE 3: Concept of a metamodel and three submodels. The basic steps of a mental simulation and anticipation are denoted within the metamodel.

2.4.1 Metamodel

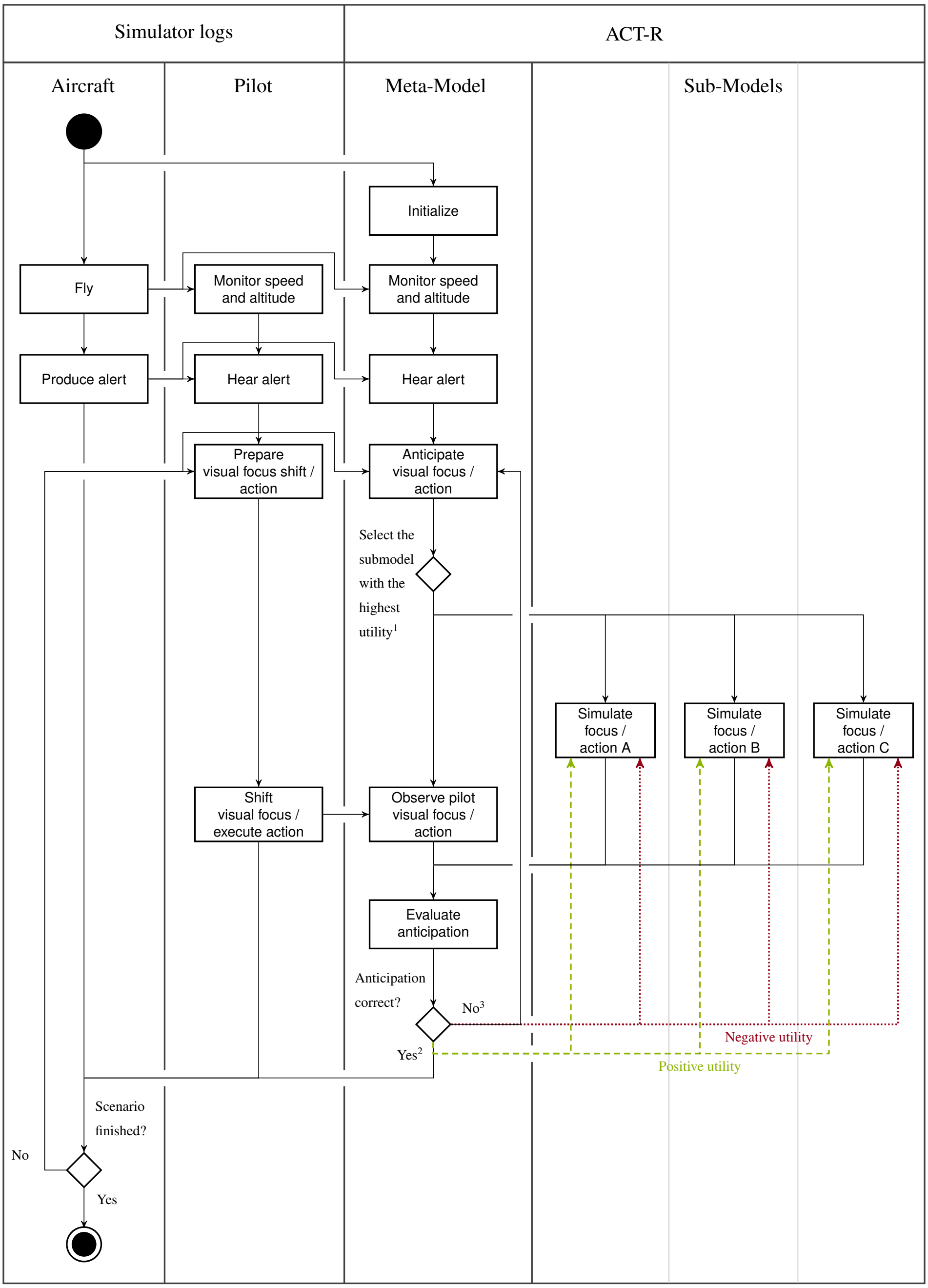
The metamodel is based on two processing stages: (1) monitoring airspeed and altitude of the aircraft, and (2) processing an alert tone and starting the mental simulation of submodels. The selection of situation models for simulation rests upon ACT-R's utility learning to incorporate learning from experience of previous anticipations. There is one production for each submodel to simulate an action, whereas the action to be simulated is specified via a chunk of the goal buffer. Utilizing only one production over multiple actions allows to include learning from experience using utility learning. A positive reward is propagated back to each production responsible for simulating a situation model when the evaluation resulted in a correct anticipation of the simulated action. If an anticipation fails, no reward is given. The simulation order of submodels is determined by their utility. The first simulated submodel is chosen by default, as all utilities have the same value initially. If the anticipation of a simulated submodel's single action (e.g., a slow response) matches the observation of a pilot's action, the metamodel continues to anticipate the next action; if not, alternative submodels will be simulated until the results match the pilot's action. If no simulated action matches the observed pilot's action later than the mean slow response plus 3 standard deviations, the metamodel continues to simulate the next action. The scenario simulation is finished after the anticipation of the last action. A schema of the model can be seen in Figure 4.

2.4.2 Submodel

The simulation of submodels takes place each time the metamodel needs to anticipate a pilot's single action. When a submodel is simulated, only one action is produced as a sequence of productions and afterwards forwarded to the metamodel. Instead of reproducing response speed themselves, submodels classify observed pilot actions as slow, medium, or fast. In line with the identified clusters classification time windows of size +/-3 standard deviations around the respective cluster's mean are used. The observed pilot action is performed by the manual module pressing a specific key (e.g., "s" for slow) and forwarded to the meta model, which retrieves the respective time window from its declarative memory.

2.5 Model evaluation

The described model, henceforth referred to as the individually simulating model (ISM), is evaluated in comparison to a normative model (NM) incapable of mentally simulating individual differences. Whereas the submodels of the ISM are based on the individual differences in pilots‘ visual focus and RTs, the NM is based on one submodel depicting the average RT of pilots. Each model was simulated thirteen times for each pilot, resulting in 104 simulations per model in total to provide robustness against the order of simulations. The order of pilots has been randomized for each of the 13 iterations. The implementation of utility-learning within the ISM is further evaluated to test the effect of experience on the number of mental simulations needed for correct anticipation. Additionally, the accuracy



[1] If more than one sub-model has the highest utility, select sub-model randomly and observe pilot focus/action.

[2] If anticipation correct, continue and propagate positive utility to simulated sub-model (green dashed line).

[3] If anticipation not correct, repeat simulation and propagate negative utility to simulated sub-model (red dotted line).

FIGURE 4: Diagram depicting a schematic implementation of the ISM. The black dot represents the start and the circled black dot represents the end point. The rectangles depict activity and the diamonds decisions.

of the ISM with utility learning has been evaluated applying a leave-one-subject-out cross validation. Training was conducted on seven subjects and testing was performed on the remaining one. The cross validation was done five times for each pilot, each iteration with a randomized order of pilots in the training runs. Model performance is evaluated in four dimensions with tests of significance carried out at $\alpha = 0.05$.

(1) The accuracy of anticipation indicates whether a model could correctly anticipate the behavior of a pilot. The model simulated the participant's visual focus area at the time of the alert sound and evaluated it based on empirical data as correct or not correct. Based on a correct comparison, it predicted whether the participants were to show a slow, medium, or fast response for the actions Thrust Lever and Engine Master according to time windows described in section 2.4.2. Each prediction was evaluated with the parameter recordings of the participant's simulator session and scored with 1 if correct, and 0 if incorrect. The accuracy is quantified using a Wilcoxon signed-rank test for pairwise comparison and Cliff's Delta denoting the effect size. The accuracy of the ISM is further evaluated using a binomial test to test whether it outperforms chance level of 33%.

(2) The deviation of the anticipated RT is evaluated to further quantify the model's accuracy in anticipating a pilot's responses. The time required by a participant or the respective submodel to carry out a response is counted in temporal-ticks (e.g., 15 ticks[3] since the engine fire alert was perceived) and compared to the respective cluster's mean RT (e.g., 11 ticks for the faster MFD/ FCU cluster). For the analysis, the RT was calculated by the runtime of ACT-R to ensure a better interpretability. The deviation is quantified as the delta between the cluster's mean and the participant's RT (in seconds) and tested for significance using a Welch two-sample t-test. Cohen's D denotes the effect size.

[3] ACT-R's temporal module provides a timer which counts the number of "ticks" that have passed since the timer was started. Tick lengths' noise increases as ticks increase. In this model, the timer was started when the alert tone was perceived. Parameters were set to maximize comparability to time in seconds.

(3) The submodel selection by utility learning is quantified by the selection rank denoting how many sub-simulations were needed until a submodel's result matched an observed pilot's action (e.g., rank 2 indicates that after two simulations a submodel matched the observed pilot's action.). The selection process of submodels is only relevant in the ISM, which was simulated twice per pilot, once with utility learning ([:u t]) and once without utility learning ([:u nil]). The model without utility learning does not learn which model to select over multiple model simulations and thus each simulation starts at random chance level. A Friedman rank-sum test was used to quantify the simulated ranks; Cliff's Delta denotes the effect size.

(4) The accuracy of cross-validated utility learning was quantified by comparing the models accuracy in the actions Thrust Lever and Engine Master to simulation runs of a model without utility learning. The prediction is scored similar as in analysis (1). The accuracy of both models is compared using a Wilcoxon signed-rank test for pairwise comparison and Cliff's Delta denoting the effect size.

3 Results

3.1 Cluster results

The cluster results indicate three different groups among pilots separated by an average silhouette width of 0.7 that differ in values for the features Focus, RT Thrust Lever, and RT Engine Master (see Figure 5). Pilots who focused on the navigation display (ND) reacted slowly in both the actions, Thrust Lever and Engine Master, whereas pilots who focused on the primary flight display (PFD) showed a medium RT. Since pilots who focused on the flight control unit (FCU) and multi-function display (MFD) reacted fastest, these visual focus areas were grouped in the same cluster named MFD. A linear regression analysis to examine the pilots' experience in flight hours showed a relationship between experience and RT Thrust Lever *F(1, 6) = 8.963*, *p = 0.0242*, R^2 *= 0.53*, and between experience and Focus *F(1, 6) = 12.043*, *p = 0.01331*, R^2 *= 0.612,* but not between experience and RT Engine Master *F(1, 6) = 3.318*, *p = 0.1184*, R^2 *= 0.25.* Pilots with high experience focused on the MFD and FCU after the alert and responded generally faster than pilots with less experience.

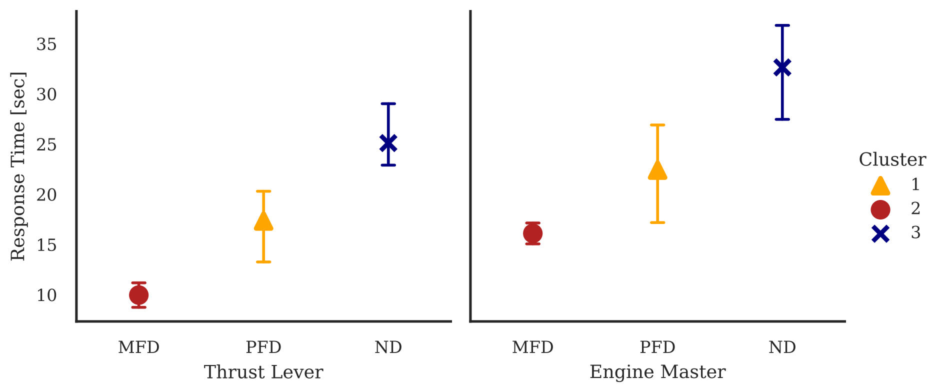


FIGURE 5: Groups of pilot behavior, based on the visual focus and the RT, separated by action (MFD = Multi-Function Display, PFD = Primary Flight Display, ND = Navigation Display; see also Figure 1 for positions).

3.2 Model Performance

(1) Accuracy of anticipation: The ISM correctly anticipated pilots' behavior for 71/104 simulations for the first action (Thrust Lever; *accuracy = 0.62*) and 49/104 for the second action (Engine Master; *accuracy = 0.47*). The NM correctly anticipated pilots' behavior for 26/104 simulations for the first action (*accuracy = 0.25*) and 37/104 for the second action (*accuracy = 0.36*). A Wilcoxon signed-rank test showed that the ISM has a significantly higher accuracy compared to the NM for the first action $V = 1560$, $p = .001$, $d = -0.43$, and for the second action $V = 630$, $p < .05$, $d = -0.115$. See also Figure 6. The binomial test showed that the ISM has a significant higher probability of success than 33% chance level for the Thrust Lever action $p < .001$, $n = 104$, $1\text{-}\beta = 0.90$ and the Engine

Master action p = .0019, n = 104, 1-ß = 1.

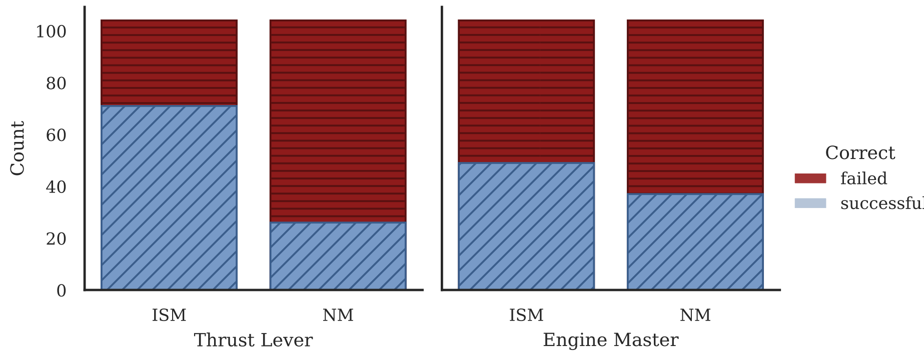


FIGURE 6: Accuracy per simulated cognitive model denoting correct and failed anticipations.

(2) Deviation of anticipated RT: A Shapiro-Wilk test for normality showed significant results, indicating the data are not distributed normally. Variance homogeneity between both model results does not exist for action one $F(74, 103) = 0.43, p < .001$ and action two $F(75, 103) = 0.40, p < .001$. A Wilcoxon rank-sum test showed a significantly smaller deviation in the ISM compared to the NM for the first action ISM: $\tilde{x} = 2.64$; NM: $\tilde{x} = 3.58$; $V = 2521, p < .001, d = 0.65$ and the second action ISM: $\tilde{x} = 3.57$; NM: $\tilde{x} = 4.315$; $V = 2366, p < .001, d = 0.93$. See Figure 7.

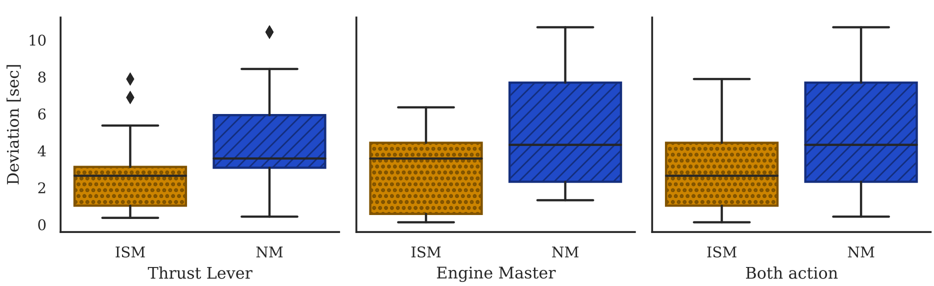


FIGURE 7: Deviation of the anticipated RT from the mean of an individual pilot group.

(3) Submodel selection by utility learning: The rankings per ISM are shown in Figure 8. Results show that more submodels matched after their initial simulation in the ISM

with utility learning compared to the ISM without utility learning. A Friedman rank-sum test showed non-significant differences between both models for the submodel selection of the Focus χ^2 = *9.03, p = .1078, d = 0.22*, the action Thrust Lever χ^2 = *8.69, p = .122, d = -0.41* and the action Engine Master χ^2 = *9, p = .1091, d = -0.08*.

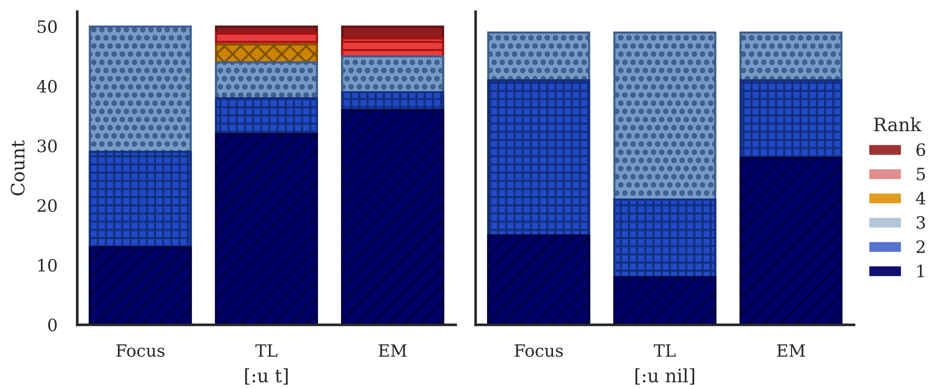


FIGURE 8: Results of the submodel selection in the ISM, denoted by the rank of how many submodels had to be simulated to match a pilot monitoring's action (TL = Thrust Lever, EM = Engine Master).

(4) Accuracy of cross-validated utility learning: The leave-one-out trained model with utility learning correctly anticipated pilots' behavior for 28/29 simulations for the first action (Thrust Lever; *accuracy = 0.97*) and 16/29 for the second action (Engine Master; *accuracy = 0.55*). The model without utility learning correctly anticipated pilots' behavior for 15/29 simulations for the first action (*accuracy = 0.52*) and 6/29 for the second action (*accuracy = 0.21*). A Wilcoxon signed-rank test showed that the utility learning model has a significantly higher accuracy compared to the model without utility learning for the first action *V = 112, p < .001, d = 0.45*, and for the second action *V = 110, p < .01, d = 0.35*. See also Figure 9.

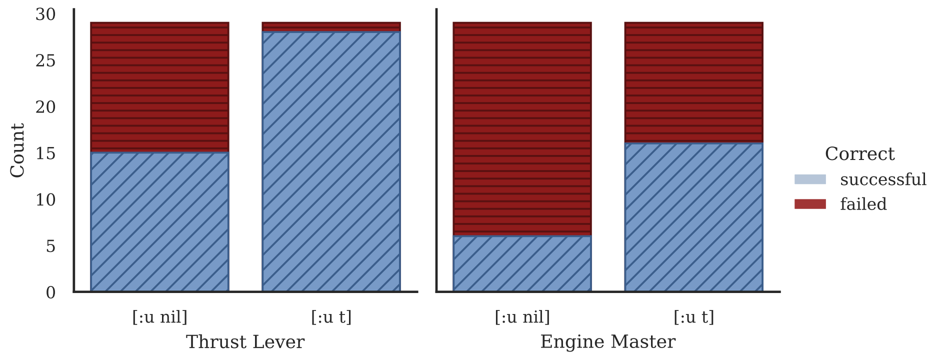


FIGURE 9: Accuracy per simulated action for a cross-validated utility learning model and a model without utility learning.

4 Discussion

The results demonstrate the feasibility of modeling anticipation of pilot actions by combining model-tracing and mental simulation.

4.1 Cluster identification and interpretation

Participating pilots were clustered into groups of similar log and eye tracking data. Assumptions about commonalities in situation models within clusters can be made based on the relationships between focus areas and reaction times. In regard to the pilots' visual focus, the MFD and FCU are closest to the WD that shows the alert message. This co-location of focus areas suggests that pilots focusing on MFD and FCU required shorter saccades to read alert messages (as opposed to head movements e.g. from the PFD). Also, alert messages popping up on the WD are perceived in peripheral vision more likely when focusing on the FCU and MFD compared to PFD and ND. As a result, these pilots would react more quickly to events. Also, ND and PFD monitoring are required for navigating and flying, so switching from demanding tasks could delay the shift of attention. Consequently, pilots looking at these displays could need more time to respond. The regression analysis showed that experience helps to make inferences about situation models. Participants with a higher number of total flight hours were assigned to cluster one, which suggests that more experienced pilots assess situations faster.

The identification of mental models depends on the availability, richness, and validity of data (Carley & Palmquist, 1992). This can be mitigated by including physiological (e.g., eye-tracking) data and grounding a mental model upon them. It could not be detected whether participants paid attention to specific sensory input or were able to make sense of it. Additional physiological measures could shed light onto differences in situation models between clusters and explain why participants focused on specific displays or showed delayed responses. For example, a combination of eye tracking and event-related electroencephalography (EEG; Klaproth, Vernaleken, et al., 2020) could provide physiological markers for situation perception and comprehension and help assess situation models in detail.

The level of detail in situation models required for effective anticipation however remains elusive. Results show that mental simulation of abstract situation models as implemented in this study can be sufficient to anticipate action. Labeling situation models within the propositional or symbolic boundaries of natural language however is delicate, as they are ad-hoc and dynamic instantiations of more generic models of systems and the environment.

4.2 Anticipation of individual pilot behavior

Individual model selection increased the accuracy in anticipating pilot actions by 37 percent in the first action and about 13 percent in the second action, resulting in a higher degree of explained variance in pilot behavior. The anticipated RT from the group mean showed a significantly lesser deviation in the ISM for both actions, which indicates more accurate quantitative anticipation of the ISM compared to the NM. In line with Klaproth, Halbrügge, et al. (2020) these findings suggest that modeling individual cognitive dynamics can improve the anticipation of an individual pilot's behavior. For example, in case of an anticipated delayed response to windshear in approach, assistance can make pilots aware of increased local windshear risk and thereby speed up their response in initiating go-arounds to reduce fuel consumption and stress on the aircraft. Anticipation of mental models will also enable cognitive assistance to engage in model reconciliation (Kambhampati, 2019), the process of adjusting inadequate mental models to better fit the situation, for example in cases of automation surprise (Sarter, Woods, & Billings, 1997). These often dangerous situations are characterized by aircraft automation behavior that was unexpected by the pilot and caused by an insensitivity of automation to pilots' mental and mental models.

Figure 6 suggests that NM anticipation accuracy increases from the first to the second action, while it decreases in the ISM. Given the small number of participants, this could be due to both models approaching chance level and the ISM's advantage decreasing with each incorrectly anticipated action. Longer sequences of action and anticipation could help testing this assumption.

Situation models were predefined and immutable which means they could not be updated with new information (Brown, Karthaus, Rehak, and Adams, 2009; Klein et al., 2007, p. 141). Future anticipation models could try to make mental models more dynamic in terms of their contents, for example to adapt strategies and solve new problems (Prezenski, Brechmann, Wolff, & Russwinkel, 2017). Also, artificially intelligent agents are neither bound to simulating a limited number (i.e., three) of models, nor to simulating them sequentially when anticipating human operators' actions. While the sequential processing model in this study is in line with the human limitations of performing multiple tasks at the same time (see Zylberberg, Ouellette, Sigman, and Roelfsema, 2012) research on primates indicates that parallel decision making does indeed occur (Zylberberg et al., 2017). Parallel simulation of a higher number of situation models may lead to conflicts with ACT-R architectural assumptions and a higher risk of overfitting in exchange for faster anticipation results and more powerful assistance.

A more refined strategy defined within mental models might be achieved by more or different types of data, such as neurophysiological data, which could contribute to understanding higher cognitive or affective states of a user (Brouwer, Zander, van Erp, Korteling, & Bronkhorst, 2015) or phenomena like inattentional deafness (Dehais, Roy, & Scannella, 2019). Anticipation accuracy of the ISM can be increased with richer data for cluster identification, as well as choosing events or situations in simulator study design that are more separable in terms of possible responses. A bigger data set could diminish the limitations imposed by this works' small sample size that requires high effect sizes to uncover significant differences (Slavin & Smith, 2009). Given the small number of data, the HCA can be considered an exploratory tool to identify potential patterns; these patterns however may not generalize well to larger samples or different situations.

4.3 Submodel selection by utility learning

The effect of utility learning on the selection process for submodel simulation and anticipation accuracy was tested. No significant differences were found between the selection rank-sums of the model with utility learning ([:u t]) and without utility learning ([:u nil]), implying that utility learning does not lead to faster selection of adequate situation models. Yet, the increase of the first rank ratio (see dark blue bars in Figure 8) in rank sums for the pilot's first action (Thrust Lever) and the second action (Engine Master) suggest an increased positive effect of utility learning on adequate submodel selection for later actions. Results of model-[:u t] show monotonous improvement with increasing number of actions, whereas no clear trend can be seen in model-[:u nil]. We assume that the more actions can be observed from an individual pilot over the course of a flight, the more quickly cognitive assistance will be able to select the most fitting mental models for anticipation with the help of utility learning.

Utility learning resulted in uneven distribution of utilities among submodel selection productions. Thanks to persevering retrieval of high-utility productions, the model with utility learning ([:u t]) needed four, five, or even six cycles of submodel simulations in some cases to select the correct submodel. Not affected by this perseverance, also known as mental set or "Einstellung" effect (Lovett & Anderson, 1996), the maximum number of simulations for the model-[:u nil] was three. Overcoming a mental set can be time-consuming and block the selection of more appropriate submodels, which can explain the lack of significant results despite the improvements and lower ranks of model-[:u t]. Cognitive assistance does not need to emulate such cognitive biases.

Submodel accuracy improved significantly with utility learning as demonstrated by cross-validation results in Figure 9. This suggests that flight recorder data like the simulator log files used in this study not only allow clustering participants, but also help a cognitive model to learn and improve its anticipation capabilities of pilot behavior. More robust validation of mental model learning would require larger sample sizes for splitting the data into training, validation and test set.

4.4 Outlook

The presented approach shows how pilot assistance could benefit from Langley's cognitive systems paradigm implemented by combining the cognitive architecture ACT-R and machine learning. It exhibits flexible capabilities like a human assistant as it selects between submodels relating to individual behavior and learns the selection process from experience. In future work further mechanisms will be investigated to better anticipate the human partner. Well designed studies should compare model performance to human anticipation processes for a better human machine collaboration. The comparison to the NM shows the added value of mental simulation for the anticipation of pilot behavior, but no statements can be made about the validity in representing human-like anticipation processes. Making mental simulation explicit and measurable can be challenging and might benefit substantially from clever task design and more controlled laboratory settings. Follow-up work could focus on implicit processes of anticipation (Poli, 2010; Riegler, 2003) by modeling situation models as declarative memory chunks. Instead of explicit selection based on separate productions per submodel, situation model selection could then be implemented as open retrieval of chunks with the highest activation (Hough, Larue, & Juvina, 2019; Thomson, Lebiere, Anderson, & Staszewski, 2015). Finally, combining machine learning and cognitive architectures could be enhanced by replacing human cluster interpretation with state of the art AI algorithms such as logic tensor networks (Serafini & Garcez, 2016).

Acknowledgements

The authors want to thank Dan Bothell, Juliane Pawlitzki and Valerie Dautry-Emery for their support.

References

Amos-Binks, A., & Dannenhauer, D. (2019). Anticipatory Thinking: A Metacognitive Capability. *ArXiv*, *abs/1906.12249*, 1-13.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775-779. doi: 10.1016/0005-1098(83)90046-8

Ball, J., Myers, C., Heiberg, A., Cooke, N. J., Matessa, M., Freiman, M., & Rodgers, S. (2010, sep). The synthetic teammate project. *Computational and mathematical organization theory*, *16*(3), 271-299. doi: 10.1007/s10588-010-9065-3

Brouwer, A.-M., Zander, T. O., van Erp, J. B. F., Korteling, J. E., & Bronkhorst, A. W. (2015, apr). Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. *Frontiers in Neuroscience*, *9*, 136. 25983676. doi: 10.3389/fnins.2015.00136

Brown, A., Karthaus, C., Rehak, L., & Adams, B. (2009, mar). The Role of Mental Models in Dynamic Decision-Making. *Human Systems Incorporated*, 83.

Butz, M. V., & Pezzulo, G. (2008). *Benefits of Anticipations in Cognitive Agents* (Vol. 5225). Berlin, Heidelberg: Springer Berlin Heidelberg.

Byrne, M. D., Kirlik, A., Fleetwood, M. D., Huss, D. G., Kosorukoff, A., Lin, R.-S., & Fick, C. S. (2004, sep). A Closed-Loop, Act-R Approach to Modeling Approach and Landing with and without Synthetic Vision System (SVS) Technology. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(17), 2111-2115. doi: 10.1177/154193120404801707

Carley, K., & Palmquist, M. (1992, mar). Extracting, representing, and analyzing mental models. *Social Forces*, *70*(3), 601. doi: 10.2307/2579746

Dehais, F., Roy, R. N., & Scannella, S. (2019, mar). Inattentional deafness to auditory alarms: Inter-individual differences, electrophysiological signature and single trial classification. *Behavioural Brain Research*, *360*, 51-59. 30508609. doi: 10.1016/j.bbr.2018.11.045

Demir, M., McNeese, N. J., Cooke, N. J., Ball, J. T., Myers, C., & Frieman, M. (2015, sep). Synthetic teammate communication and coordination with humans. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 951-955. doi: 10.1177/1541931215591275

Endsley, M. R. (1995, mar). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32-64. doi: 10.1518/001872095779049543

Endsley, M. R. (2000, jul). Situation models: an avenue to the modeling of mental models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*(1), 61-64. doi: 10.1177/154193120004400117

Estes, S., Burns, K., Helleberg, J., Long, K., Stein, J., & Pollack, M. (2016). Digital Copilot: Cognitive Assistance for Pilots. In (p. 141-144). AAAI Press.

Flumeri, G. D., Crescenzio, F. D., Berberian, B., Ohneiser, O., Kramer, J., Aricò, P., … Piastra, S. (2019, sep). Brain-Computer Interface-Based Adaptive Automation to Prevent Out-Of-The-Loop Phenomenon in Air Traffic Controllers Dealing With Highly Automated Systems. *Frontiers in Human Neuroscience*, *13*, 296. 31555113. doi: 10.3389/fnhum.2019.00296

Fu, W.-T., Bothell, D., Douglass, S., Haimson, C., Sohn, M.-H., & Anderson, J. (2006). Toward a real-time model-based training system. *Interact. Comput.*, *18*(6), 1215-1241.

Furlough, C. S., & Gillan, D. J. (2018, dec). Mental models: structural differences and the role of experience. *Journal of cognitive engineering and decision making*, *12*(4), 269-287. doi: 10.1177/1555343418773236

Halbruegge, M. (2013). *ACT-CV - Bridging the gap between cognitive models and the outer world.* Berlin: TU Berlin.

Hegarty, M., Just, M. A., & Morrison, I. R. (1988, apr). Mental models of mechanical systems: individual differences in qualitative and quantitative reasoning. *Cognitive Psychology*, *20*(2), 191-236. 3365939. doi: 10.1016/0010-0285(88)90019-9

Hough, A. R., Larue, O., & Juvina, I. (2019). A Metacognitive Triggering Mechanism for Anticipatory Thinking. *Unpublished*. doi: 10.13140/rg.2.2.19337.80483

Kambhampati, S. (2019). *Challenges of Human-Aware AI Systems.*

Kennedy, W. G., Bugajska, M. D., Adams, W., Schultz, A. C., & Trafton, J. G. (2008). Incorporating Mental Simulation for a More Effective Robotic Teammate. *Pro-*

ceedings of the 23rd National Conference on Artificial Intelligence, *3*, 1300-1305.

Klaproth, O. W., Halbrügge, M., Krol, L. R., Vernaleken, C., Zander, T. O., & Russwinkel, N. (2020). A neuroadaptive cognitive model for dealing with uncertainty in tracing pilots' cognitive state. *Topics in cognitive science*, *12*(3), 1012-1029. 32666616. doi: 10.1111/tops.12515

Klaproth, O. W., Vernaleken, C., Krol, L. R., Halbruegge, M., Zander, T. O., Russwinkel, N., et al. (2020). Tracing pilots' situation assessment by neuroadaptive cognitive modeling. *Frontiers in Neuroscience*, *14*, 795. 32848566. doi: 10.3389/fnins.2020.00795

Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). A data-frame theory of sensemaking. *Expertise out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making*, 113-155.

Kotseruba, I., & Tsotsos, J. K. (2018, jul). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, 1-78. doi: 10.1007/s10462-018-9646-y

Langley, P. (2006). Cognitive Architectures and General Intelligent Systems. *{AI} Magazine Volume*, *27*(2), 33-44.

Langley, P. (2011, mar). The changing science of machine learning. *Machine learning*, *82*(3), 275-279. doi: 10.1007/s10994-011-5242-y

Langley, P. (2012). The Cognitive Systems Paradigm. *Advances in Cognitive Systems*, *1*, 3-13.

Lovett, M. C., & Anderson, J. R. (1996, oct). History of success and current context in problem solving. *Cognitive Psychology*, *31*(2), 168-217. doi: 10.1006/cogp.1996.0016

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming With a Synthetic Teammate: Insights into Human-Autonomy Teaming. *Human Factors*, *60*(2), 262-273. 29185818. doi: 10.1177/0018720817743223

Pandey, A. K., Ali, M., & Alami, R. (2013, apr). Towards a Task-Aware Proactive Sociable Robot Based on Multi-state Perspective-Taking. *International journal of social robotics*, *5*(2), 215-236. doi: 10.1007/s12369-013-0181-3

Pezzulo, G., Butz, M. V., & Castelfranchi, C. (2008). *The Anticipatory Approach: Definitions and Taxonomies* (Vol. 5225). Berlin, Heidelberg: Springer Berlin Heidelberg.

Poli, R. (2010, jun). The many aspects of anticipation. *Foresight (Los Angeles, Calif.)*, *12*(3), 7-17. doi: 10.1108/14636681011049839

Prezenski, S., Brechmann, A., Wolff, S., & Russwinkel, N. (2017, aug). A cognitive modeling approach to strategy formation in dynamic decision making. *Frontiers in Psychology*, *8*, 1335. 28824512. doi: 10.3389/fpsyg.2017.01335

Riegler, A. (2003). *Whose Anticipations?* Springer Verlag.

Russwinkel, N., Vernaleken, C., & Klaproth, O. W. (2020). *Towards cognitive assistance and teaming in aviation by inferring pilots mental state* (Vol. 1131). Cham: Springer International Publishing. doi: 10.1007/978-3-030-39512-4_155

Salas, E. (2017). *Situational Awareness*. CRC Press.

Sarter, N., Woods, D., & Billings, C. (1997). *Automation surprises* (Vol. 2).

Scharfe, M., & Russwinkel, N. (2019). Towards a Cognitive Model of the Takeover in Highly Automated Driving for the Improvement of Human Machine Interaction..

Schmidt, J., Braunagel, C., Stolzmann, W., & Karrer-Gauss, K. (2016, jun). Driver drowsiness and behavior detection in prolonged conditionally automated drives. In *2016 {IEEE} Intelligent Vehicles Symposium ({IV})* (p. 400-405). IEEE. doi: 10.1109/IVS.2016.7535417

Schoppek, W., & Boehm-Davis, D. A. (2004). Opportunities and challenges of modeling user behavior in complex real world tasks. *{MMI} interaktiv*, *7*, 47-60.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006, feb). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70-76. 16406326. doi: 10.1016/j.tics.2005.12.009

Serafini, L., & Garcez, A. S. D. (2016). Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge. *CoRR*, *abs/1606.04422*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016, jan). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484-489. 26819042. doi: 10.1038/nature16961

Slavin, R., & Smith, D. (2009). The Relationship Between Sample Sizes and Effect Sizes in Systematic Reviews in Education. *Educ. Eval. Policy Anal.*, *31*(4), 500-506.

Somers, S., & West, R. (2013, jul). Steering Control in a Flight Simulator Using ACT-R. In *Proceedings of the International Conference on Cognitive Modeling* (p. 233-238).

Steiger, D. M., & Steiger, N. M. (2009). Discovering a Decision MakerMental Model with Instance-Based Cognitive Mining:. *Interdisciplinary Journal of Information, Knowledge, and Management*, *4*, 001-022. doi: 10.28945/61

Taatgen, N. A., Huss, D., Dickison, D., & Anderson, J. R. (2008, aug). The acquisition of robust and flexible cognitive skills. *Journal of Experimental Psychology: General*, *137*(3), 548-565. 18729715. doi: 10.1037/0096-3445.137.3.548

Teo, G., Matthews, G., Reinerman-Jones, L., & Barber, D. (2020, dec). Adaptive aiding with an individualized workload model based on psychophysiological measures. *Human-Intelligent Systems Integration*, *2*(1-4), 1-15. doi: 10.1007/s42454-019-00005-8

Thomson, R., Lebiere, C., Anderson, J. R., & Staszewski, J. (2015, sep). A general instance-based learning framework for studying intuitive decision-making in a cognitive architecture. *Journal of Applied Research in Memory and Cognition*, *4*(3), 180-190. doi: 10.1016/j.jarmac.2014.06.002

Thomson, R., Lebiere, C., & Bennati, S. (2014, apr). Human, model and machine: A complementary approach to big data. In *Proceedings of the 2014 Workshop on Human Centered Big Data Research - {HCBDR} '14* (p. 27-31). New York, New York, USA: ACM Press. doi: 10.1145/2609876.2609883

Vinokurov, Y., Lebiere, C., Wyatte, D., Herd, S., & O'Reilly, R. (2012, jan). Unsurpervised learning in hybrid cognitive architectures. *{AAAI} Workshop - Technical Report*, 36-41.

Zylberberg, A., Lorteije, J. A., Ouellette, B. G., Zeeuw, C. I. D., Sigman, M., & Roelfsema, P. (2017, jun). Serial, parallel and hierarchical decision making in primates. *eLife*, *6*. 28648172. doi: 10.7554/eLife.17331

Zylberberg, A., Ouellette, B., Sigman, M., & Roelfsema, P. R. (2012, oct). Decision making during the psychological refractory period. *Current Biology*, *22*(19), 1795-1799. 22921368. doi: 10.1016/j.cub.2012.07.043