A Dual-Process Cognitive Model for Testing Resilient Control Systems

Jim Blythe
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
Email: blythe@isi.edu

***Abstract*—We describe an agent-based model of individual human behavior that combines a dual-process architecture with reactive planning and mental models in order to capture a wide range of human behavior, including both behavioral and conceptual errors. Human operator behavior is an important factor in resilient control of systems that has received relatively little attention. Models of human behavior and decision making are needed in order to test existing control systems under a range of conditions or analyze possible new approaches. While the model we describe has been developed and applied in the area of cyber security, it is relevant to a wide range of resilient control systems that include human operation. We discuss an application to modeling operator behavior in a nuclear power plant.**

I. Introduction

Human operator behavior is an important but relatively under-studied factor in resilient control.

Human Reliability Analysis (HRA) incorporates human behavior as part of a probabilistic risk assessment, but most (HRA) techniques have performed a static analysis that does not capture variations in human behavior, or alternative potential outcomes from an initial situation. Recently, however, there has been work in simulations that can capture dynamic effects and individual differences given an appropriate model of human behavior and of the domain [1], [2].

The success of this approach depends on the accuracy of models of human decision making under different circumstances, taking into account the information available to the decision maker, their understanding of the control system, the time available to make a decision and individual differences between decision makers. In this paper we present a model of cognitive behavior that can be used as part of such an approach, and describe implemented software agents that use this model. The model is under development with an emphasis on cyber security applications in the Deter project [3], [4]. Our agents use a belief-desire-intention (BDI) framework to model deliberative but responsive action, combined with a set of associative rules, that model human behavioral actions, and mental models, that model human conceptual reasoning that may be inaccurate.

We provide an overview of our agent model and compare the approach with other cognitive modeling platforms, such as ACT-R and SOAR. The agent system is planned to be made available through the Deter platform.

A cognitive architecture that is adequate for modeling human operators should provide the capability to combine deliberative action with responsiveness to unexpected changes in the environment. In order to model human errors of different kinds, the architecture should also capture dual-process models that combine intuitive and rational actions, as well as a succinct representation for human mental models that may lead to suboptimal behavior that is rational according to the agent's beliefs.

Our platform satisfies all these criteria. The approach combines a fast recognize-suggest process, that continually suggests both actions and related concepts based on stimuli, with a slower, rational process that can override the suggestions of the fast process using abstract reasoning [5]. This reasoning follows the approach of reactive planning to continually re-evaluate the agent's goals based on the environment and the fast recognition process, so that plans may be modified or abandoned as the situation changes [6]. The agent's approach for evaluating goals and candidate plans is based on internally executing mental models [7], that capture the operator's beliefs about the domain, which may be incorrect.

This framework can model human reasoning at a variety of timescales, including fast reactive behavior and slower deliberative reasoning as well as long-term strategic reasoning. It can be used both to model best practices for operator control as well as to investigate the impact of misconceptions and boundedly rational behavior. After introducing the framework, we demonstrate these modes of behavior in a scenario involving control of a nuclear power plant.

II. Related work

The cognitive architecture explored in this paper is designed to be a component of a software testing suite that may include software simulations of the system under control and possibly physical and physiological models of human activity. We begin by considering cognitive components and then turn to related work embedding these components in broad simulations.

A large body of work exists on computer architectures for modeling human information processing, among the most notable being SOAR [8], ACT-R [9] and Icarus [10]. SOAR combines a universal subgoaling approach for reasoning with a universal learning mechanism. Given a problem, its set of forward-chaining production rules add new information into

978-1-4673-0162-6/12/$31.00 ©2012 IEEE

working memory in an attempt to find a solution. If an impasse is reached, SOAR will automatically form a new problem space whose purpose is to solve the impasse, a process called 'subgoaling'. New productions may fire in this subspace and, if a solution is found, the elements of the solution will be encapsulated in a new production rule that can work in the original space, a learning process called 'chunking'. ACT-R [9] is a general problem solver that makes use of an associative memory, following ideas from cognitive science. In an associative memory, nodes, representing beliefs about the environment, have activation levels that signify the current attention being paid to them. If a node receives a high activation level, neighboring nodes, representing related concepts, will receive increments in a process called 'spreading activation'. Icarus [10] is a cognitive architecture that pays particular attention to physical embodiment, for example including muscle learning as one of the learning mechanisms.

Each of these architectures explore a number of features that are believed to be important in modeling human information processing. As we describe below, we have developed a different approach in order to capture recent results in dual-process theories of cognition [5]. Dual-process theories seem well-suited to modeling human performance that is sometimes suboptimal and depends on factors such as time available, fatigue and attention levels. Much of the work in dual-process approaches came to prominence after the cognitive architectures we have described were originally designed, though other aspects of our approach are consistent with them.

Cognitive agents have been combined with simulations in immersive environments for training for many years, *e.g.* [11], [12]. These combine physical simulation with cognitive models and typically simulated human bodies. The aim is typically a believable experience for user engagement [13]. Cognitive simulation within simulated worlds have also been developed for strategy testing, *e.g.* [14], [15], but they usually do not model the human body. This combination has recently been proposed for testing control systems by Tran et al [1], who propose a plug-in architecture where any of a number of cognitive simulation modules could be used.

III. The Detergent Cognitive Model

In this section we provide details of the Deter Agent model, called the Detergent model, that is under development as an agent-based model for human behavior in the Deter cyber security platform [4]. The model follows a dual-process approach, with analytic processing provided by a reactive planner that makes use of agent mental models of the domain, that may be inaccurate and vary between agents. We now describe each of these aspects of the approach in more detail. Figure 1 shows the overall architecture of the platform.

A. Dual-process model of cognition

A point of departure of our agent approach from most existing cognitive architectures is the central position given to a dual-process model, in which two separate processes are continually running and occasionally compete for control of the agent. We follow the literature in referring to them as "System 1" and "System 2" [5], [16]. System 1 processes are rapid, automatic and largely opaque to the reasoner, creating associations and suggesting actions immediately environmental stimuli are received. System 2 is slow and sequential, allowing the agent to perform abstract reasoning but requiring limited working memory to function. Much of the time, human behavior is driven by System 1, which takes no noticeable mental effort. However System 2 can suppress System 1 activity in order to create courses of action or provide judgments on choices with a more rational basis.

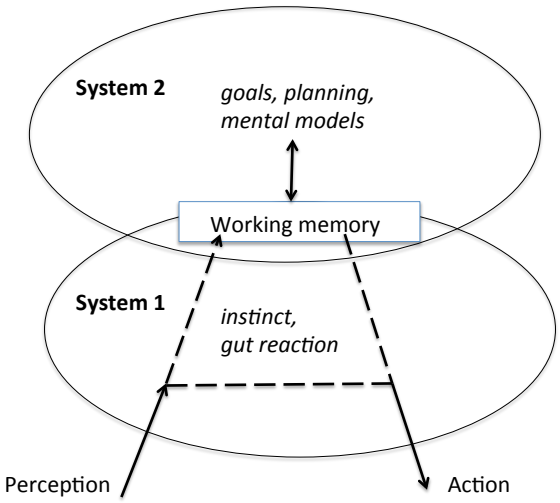

Fig. 1. Overview of the interaction between the two processes within the percept-act loop of the agent. System 2's understanding of the world depends on System 1's interpretation, which populates a finite working memory.

Evidence for a dual-process model of human behavior can be found in a variety of situations. For example, human logical reasoning varies in speed and accuracy when the semantic content of the facts and rules varies [17]. In one study, over 90% of college students reasoned correctly about syllogisms when the conclusion's truth aligned with with its commonsense believability, but a majority made mistakes when they were misaligned. This effect is called *belief-bias*. fMRI studies show that correct, logical decisions are associated with the right inferior prefrontal cortex, and incorrect, belief-based decisions with the ventral medial prefrontal cortex [18]. In examples like this, System 1 and System 2 essentially compete to provide an answer. System 1 provides an answer rapidly and relatively effortlessly based on the content of the problem, while System 2 can perform the abstract logical reasoning required at the cost of noticeable mental effort.

Our agent platform includes an explicit dual-process model, using production rules for System 1 and explicit planning with mental models for System 2 as described below. In our approach, a number of factors determine whether System 2 is active in a given situation and which suggestion is used. These include an innate relative strength of the systems, fatigue levels

in System 2, emotions of different kinds that may boost either system, and the relative confidence of the systems in their answers.

Classical models of cognition such as SOAR and ACT-R have difficulty modeling belief-bias behavior. In principle, SOAR productions at the domain level might be used to model System 1 behavior, while a more abstract problem space could be used to model System 2. However the built-in chunking mechanism will only learn production rules that correctly summarize the results of more abstract reasoning, so in practice the conflict inherent in belief-bias would not occur through this mechanism. This is generally a good thing, allowing SOAR to model the improvement in performance seen as humans learn tasks. However belief-bias and related effects are essential in modeling human operator performance in many tasks.

The dual-process model provides an effective computational architecture for models of human cognition. The interaction between System 1 and System 2 can elegantly account for differences in behavior when operators are tired, either physically or through sustained mental effort with System 2, or are under time pressure or heightened emotions. The intuition of experts can be modeled with the trained improvement of System 1 decisions and accounts for an accentuated difference under time pressure.

B. Reactive planning

In normal operation, an agent's System 1 will suggest the appropriate action for a given situation, which the agent will adopt with minimal processing from System 2. However, System 2 may become more active under a number of circumstances, including when System 1's processing indicates a low confidence in its suggestion, or if it has produced contradictory suggestions. Our agent's System 2 module primarily performs explicit planning, in a BDI-based reactive planning style [6] similar to that used in PRS [19] and SPARK [20]. This is a relatively common approach for explicitly reasoning from goals to chosen actions that allows the agent to react when the situation changes during the planning process [14].

The agent begins by computing a set of goals based on its initial environment that it will actively try to achieve. The agent then chooses from a set of methods in its library that are able to achieve each goal. The body of each method is a small script that may include subgoals that are to be met in running the method. For each method that is chosen, the agent adopts the subgoals and recursively chooses methods to achieve them. Planning generally finishes when the agent finds a complete plan, whose actions are all directly executable. The agent then chooses one action that is immediately applicable and executes it.

In order to react to changes in the environment, after executing this first action and gathering new data from the environment, the agent recomputes its goals and replans from scratch rather than re-using its existing plan. If the world is unchanged, the same plan will probably be chosen, but if there are significant changes, the agent may pick different goals or different ways to achieve them.

C. Mental models

The final component of our agent platform concerns the agent's beliefs about its environment and how it reasons with those beliefs in order to choose one action over another. In modeling human behavior, the state of the operator's beliefs about the domain is as important a factor as the cognitive machinery that is used to reason about it. Within System 2, we follow the mental models approach of Johnson-Laird and others [21], [7], which holds that we construct symbolic models of our domain to reason about it. When reasoning about simple physical domains these models typically match the structure of the domain but when they are applied to more complex domains such as a distributed control environment, the models are frequently incomplete and incorrect. However, an incomplete, simplified model may be more effective for a reasoner than a complete one, to the extent that it allows the reasoner to make good, timely decisions about its environment.

Mental models are implemented in our agent platform in terms of (1) inferences about the current state from observations and (2) predictions about future states based on possible operator actions. They are used to choose between two courses of action by comparing the expected end states of each. We discuss an application to cyber security modeling in [4].

IV. Example scenario: Three mile island

We illustrate some of the capabilities of this agent with examples based on the case of the Three Mile Island nuclear power plant. This is a well-studied case that culminated in a small leak of radioactive material in 1979. The accident might have been averted had the human operators behaved differently, although there were many independent contributing factors [22], [23].

A. Overview of events

Three Mile Island is a water-cooled reactor with two coupling circuits to remove heat from the reactor core. Coolant flows in an internal circuit around the core, exchanging heat with water that flows in an external circuit, driving steam turbines. During maintenance on the external circuit, a leaky valve indirectly caused the main pumps to close. Emergency pumps intended to maintain water flow were mistakenly left blocked after earlier maintenance. This meant that heat was not removed from the internal circuit and the reactor core began to overheat. The reactor was automatically closed down, but decay heat led to a pressure increase in the coolant, causing a relief valve to release some coolant. However, the relief valve failed to close, leading to significant loss of coolant. To avoid core melt-down, high pressure injection (HPI) pumps forced water into the internal circuit.

Operators were concerned about over-pressurization of the internal circuit and throttled the HPI pumps. Over two hours later, with a new shift of operators, the loss of coolant was recognized and rectified, but by this time there had been a

partial melt-down of the core. In theory, operators might have reasoned from indicators of the temperature that there may be a loss of coolant and that the indicator showing the relief valve was switched to closed may not mean that the valve was actually closed. In practice, operators were required to attend to many alarms in a highly time-dependent situation, making rational inference from the information available very difficult. This was not a one-off case: in an earlier incident at a different power plant the relief valve had malfunctioned in an open position and the operators had also throttled the HPI pumps in concern over high pressure. There, disaster was averted because the mistake was noticed within twenty minutes and the plant was operating at only 9% of capacity.

B. Factors in modeling operator decisions

Ideally a simulation that included physical aspects of the plant, cognitive agents and a model of the information available through the HCI system might uncover failure modes such as this one and others. We can compare the likely responses of different cognitive agents modeling operators in this situation.

Whatever architecture is used, a rational system, in possession of all the relevant sensor information and a correct model, would be expected to infer that there was a coolant loss and avoid throttling the HPI pumps. We assume that both temperature and pressure signals are attended to and are seen as supporting the opposing failure modes of coolant loss and over-pressurization respectively. However the high pressure is consistent with coolant loss since the coolant temperature is above boiling point, while over-pressurization cannot explain the high temperature. Some models might require the relief valve (PORV) to be open in order to explain the coolant loss, which is consistent with the sensor readings given that the PORV signal indicates that the relay has attempted to close the valve, not that it is known to be closed.

A simple model of bounded attention would lead to the same behavior, assuming internal coolant temperature is attended to before the pressure and PORV readings, since it is associated with the greater potential danger.

The behavior observed at TMI could be duplicated by rational agent models that use incorrect mental models, for example models that misread the PORV indicator as reliably showing the relief valve is closed. In this approach, however, the simulated operators would be aware of the temperature readings but would choose to discount them based on their model. However, human operators may not have been aware of the temperature readings as they throttled the HPI pumps. This implies not taking steps to check these readings, which is formally irrational given the costs of checking and of misdiagnosis, and so a model of bounded rationality is required to capture the behavior. As a high-level explanation, *confirmation bias* can account for this. After operators saw the PORV and HPI signals, they may have formed a hypothesis quickly that centered on over-pressurization. This would have been strengthened by earlier training that placed more emphasis on over-pressurization than coolant loss, perhaps because it was seen as a more likely contingency. With a confirmation bias, subsequent checks of the readings by the operators would have focused on confirming this hypothesis and confirming that their actions were having a desired effect.

A dual-process account provides a mechanism that explains this behavior, and also predicts when it is likely to be more pronounced. In this account, the agent's system 1 processes the PORV, HPI and pressure readings, seeking not only a consistent view of the world but also a satisfying course of action. Based perhaps on previous training, the over-pressurization theory and throttling course of action is rapidly suggested. System 1's desire for consistency and control would lead it to ignore the temperature reading even if it were observed by the operator. The reading is therefore not presented to System 2 via the shared working memory, and so its rational model, if used to override System 1's recommendation, would not have access to the sensor data required to override it.

In our simulation, System 1's misdiagnosis is modeled by a simplistic fast reaction to the HPI and pressure readings. Spreading activation [24] can also explain the incorrect decision even when System 1 processing can take into account the effects of the coolant reaching boiling point. This is because this reasoning requires an intermediate step, going from the temperature reading to the idea of the coolant boiling. In the time this takes, the over-pressurization theory will have already been activated in memory and will essentially suppress the more nuanced reasoning about temperature, which opposes it.

Given the resources and equipped with mechanisms to reason about hypothetical information, System 2 could infer that the missing temperature reading could defeat its current hypothesis about over-pressurization, and take deliberate steps to take the reading. This level of resources are unlikely to be available to System 2 when the operator is tired (the accident took place at 4am) or under time pressure. Situations associated with high emotion will also reduce the likelihood that System 2 is engaged to double-check System 1's recommended action.

V. Discussion

The Detergent dual-process architecture has been implemented and used to model observations of human behavior in cyber-security phishing scenarios, as described in [4]. In that study, mental models derived from user studies were used to explain observed patterns of behavior (and in some cases the lack of them). Here, as we described, a combination of mental models and human biases that are related to a dual-process foundation may provide insight into operator behavior. However the nuclear plant scenario described above has not yet been fully implemented.

If potential human errors can be classified in terms of System 1 and System 2 processing along with mental models, this can provide useful directions for how to mitigate the effects with training and improved processes and working environments. Milkman et al [25] discuss how to improve decision-making, taking a dual-process view and considering a number of domain-independent strategies, including moving from System 1 to System 2 thinking when appropriate and

leveraging System 1 to improve decision making when it is likely to be dominant. In the context of this paper, removing time pressure and managing fatigue and attention levels are ways to increase the propensity for System 2 thinking, while training that emphasizes situation-based procedures in simulated accidents improves the choices made through System 1. For example, the 30-minute rule commonly used in nuclear plants today can be seen as improving the chance for System 2 to be engaged. The discussion in the last session also relied on an incorrect decision from System 1, which might be corrected with training.

Working through explicit scenarios can also highlight the relative importance of aspects of the interface, such as the availability of temperature measurements alongside pressure measurements and indicators for the HPI and PORV valves, and can be valuable in arranging the environment to suit good operator decision-making.

A. Models of Emotion

In addition to managing the effects of fatigue, time pressure and attention levels on operator decision-making, another important factor that has not been explored in this paper is emotion. Although earlier research viewed emotion as an impediment to clear decision making, recent research has shown that the emotion mechanism is crucial to the human decision process and non-neutral moods can be beneficial in some circumstances [26]. With our colleagues we are developing an architecture that models both the effects of emotion on cognitive activities such as judgment and decision making, and also how emotions evolve in response to the environment. The architecture being developed, EmoCog, is broadly compatible with the dual-process model described here, with emotion processing largely taking place in a System 1 module and affecting rational System 2 thought through similar mechanisms [27].

B. Conclusion

As we have shown, a dual-process model of human cognition such as this one can be used to verify a set of conditions that might lead to errors of commission of the kind that occurred at TMI. However there may be many such conditions, indicating different possible causes for the errors. For each of these, a model such as this might be used to find the most promising ways to reduce the probability of future errors, for example with changes to the operator interface, or with training to improve rapid decision making that may stem from System 1, or with methods to reduce the likelihood of a weakened System 2. Clearly there is much to be done to improve our current model based both on the existing literature and future empirical work. We believe that this is a very promising avenue for continuing research with potential to improve resilience in control systems by modeling human decision-making and improving its effectiveness.

References

[1] T. Q. Tran, D. I. Gertman, D. D. Dudenhoeffer, R. L. Boring, and A. R. Mecham, "Cognitive virtualization: combining cognitive models and virtual environments," in *Human Factors and Power Plants and HPRCT 13th Annual Meeting, 2007 IEEE 8th*, aug. 2007, pp. 195 –200.

[2] D. L. Kelly, R. L. Boring, A. Mosleh, and C. Smidts, "Science-based simulation model of human performance for human reliability analysis," in *Enlarged Halden Program Group (EHPG) Meeting*, 2011.

[3] T. Benzel, "The science of cyber security experimentation: The deter project," in *Annual Computer Security Applications Conference*, 2011.

[4] J. Blythe and J. Camp, "Implementing mental models," in *Workshop on Semantic Computing and Security*, 2012.

[5] K. Stanovich and R. West, "Individual differences in reasoning: Implications for the rationality debate?" *Behavioral and Brain Sciences*, 2000.

[6] M. Bratman, *Intention, plans, and practical reason*. Harvard University Press, 1987.

[7] D. Gentner and A. Stevens, *Mental Models*. Lawrence Erlbaum Associates, Inc., 1983.

[8] J. E. Laird, A. Newell, and P. S. Rosenbloom, "Soar: An architecture for general intelligence," *Artificial Intelligence*, vol. 33, no. 1, pp. 1 – 64, 1987.

[9] J. Anderson, *Rules of the Mind*. Lawrence Erlbaum Associates, Inc., 1993.

[10] P. Langley, K. B. McKusick, J. A. Allen, W. F. Iba, and K. Thompson, "A design for the icarus architecture," *SIGART Bull.*, vol. 2, no. 4, pp. 104–109, Jul. 1991.

[11] W. L. Johnson, J. W. Rickel, and J. C. Lester, "Animated pedagogical agents: Face-to-face interaction in interactive learning environments," *International Journal of Artificial Intelligence in Education*, vol. 11, pp. 47–78, 2000.

[12] J. Rickel, S. Marsella, J. Gratch, R. Hill, D. Traum, and W. Swartout, "Toward a new generation of virtual humans for interactive experiences," *Intelligent Systems, IEEE*, vol. 17, no. 4, pp. 32–38, 2002.

[13] J. Bates, B. Loyall, and S. Reilly, "Integrating reactivity, goals and emotion in a broad agent," in *Proc. Cognitive Science*, 1992.

[14] M. Tambe, W. L. Johnson, R. M. Jones, F. Kossd, J. E. Laird, P. S. Rosenbloom, and K. Schwamb, "Intelligent agents for interactive simulation environments," *AI Magazine*, vol. 16, no. 1, 1995.

[15] H. Liu and D. Nau, "Introduction to the acm tist special issue: Ai in social computing and cultural modeling," *ACM Transactions on Intelligent Systems and Technology*, vol. 1, no. 1, 2010.

[16] D. Kahneman, *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[17] J. S. Evans, "In two minds: dual-process accounts of reasoning," *Trends in Cognitive Sciences*, vol. 7, no. 10, pp. 454 – 459, 2003.

[18] V. Goel and R. J. Dolan, "Explaining modulation of reasoning by belief," *Cognition*, vol. 87, no. 1, pp. B11 – B22, 2003.

[19] M. Georgeff and A. Lansky, "Reactive reasoning and planning," in *Proc. American Association for Artificial Intelligence*, 1987.

[20] D. Morley and K. Myers, "The spark agent framework," in *Autonomous Agents and Multi-agent Systems*, 2004.

[21] P. Johnson-Laird, *Mental Models*. Harvard University Press, 1986.

[22] C. Perrow, *Normal Accidents: Living with High-Risk Technologies*. Basic Books, 1984.

[23] A. Hopkins, "Was three mile island a normal accident?" *Journal of Contingencies and Crisis Management*, vol. 9, no. 2, pp. 65–72, 2001.

[24] A. Collins and E. Loftus, "A spreading-activation theory of semantic processing," *Psychological Review*, vol. 82, no. 6, pp. 407–428, 1975.

[25] K. Milkman, D. Chugh, and M. Bazerman, "How can decision making be improved?" *Perspectives on Psychological Science*, vol. 4, no. 4, pp. 379–383, 2009.

[26] A. Bechara, H. Damasio, and A. Damasio, "Emotion, decision making and the orbitofrontal cortex." *Cerebral cortex*, 2000.

[27] J. Lin, M. Spraragen, J. Blythe, and M. Zyda, "Emocog: Computational integration of emotion and cognitive architecture," in *Proc. FLAIRS*, 2011.