Implementing Mental Models

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

L. Jean Camp
School of Informatics
Indiana University
Bloomington, IN, USA
ljcamp@indiana.edu

***Abstract*—Users' mental models of security, though possibly incorrect, embody patterns of reasoning about security that lead to systematic behaviors across tasks and may be shared across populations of users. Researchers have identified widely held mental models of security, usually with the purpose of improving communications and warnings about vulnerabilities. Here, we implement previously identified models in order to explore their use for predicting user behavior. We describe a general approach for implementing the models in agents that simulate human behavior within a network security test bed, and show that the implementations produce behaviors similar to those of users who hold them. The approach is relatively simple for researchers to implement new models within the agent platform to experiment with their effects in a multi-agent setting.**

***Keywords*-computer security; cognitive science; mental models**

I. Introduction

A good experimental infrastructure is critical to developing effective new cyber security technology [1]. Many if not most attacks rely on human action, for example phishing attacks or those that rely on misconfigured security, and so the experimental infrastructure must be able to reflect the impact of human activity on the systems under test. While human behavior is not as predictable as that of, say, a router box or a computer running linux, there are patterns of behavior that allow probabilistic prediction, particularly over groups of people or long-term interactions.

Perhaps the main predictor of human security behavior is the level and structure of the individual's knowledge about security. At the simplest level, users cannot look for attacks or apply security tools they are not aware of. Furthermore, the motivation to inspect messages and web sites or apply tools depends on the user's belief about their susceptibility to an attack, its potential severity and the cost and efficacy of preventive or mitigating behavior.

Several researchers have investigated the knowledge that non-experts have about security within the framework of *mental models*. These are internal models that humans use to reason about the world, widely studied in cognitive science. Researchers in security have elicited mental models with the aim of improving communication with users, improving education about security or improving the interfaces of security tools. In this paper, we consider these models as the basis of an agent model of human behavior that can be used in a security test bed, such as DETER [1] or the National Cyber Range [2]. Mental models can increase the predictive power of these agents when there are commonalities in the models used within a group, or when models used by one individual lead to a pattern of behavior across several tasks. There is evidence for both conditions in the experiments that we summarize below.

We describe implementations of models collected from non-experts and show that they can reproduce observed security behavior. Our declarative implementations can be viewed as semantic models of human security-related beliefs and behavior. They adhere to the typical approach of mental models, *e.g.* Gentner and Stevens [3], in that the models support internal simulations that are used to decide on security behavior. We describe the implementation of a set of models elicited by Wash [4] and show that the agents make broadly similar decisions to those reported by human subjects. Finally we discuss future directions, including the use of analogy with mental models that are not directly concerned with security and recognizing user's models for adaptive interfaces.

II. Mental models

It has long been held that humans reason about their world by manipulating internal, symbolic models [5], [3], [6]. When reasoning about simple physical domains these models typically match the structure of the domain and humans reason about future events through simulation. When they are applied to more complex domains, such as when making decisions about medical treatments or computer security, these models are more likely to be incorrect or incomplete, bearing a looser relation to the structure of the real-world situation, much of which may be unknown to the human reasoner. The models can still be effective, however, to the extent that they allow the reasoner to make better decisions than would be possible without them.

Mental models in computer security have been studied for two distinct reasons. First, to build more effective interfaces by understanding the users' model of security [7], [8] and second as a tool for effective communication

DOI 10.1109/SPW.2012.31

with users [9], [10], [11]. In this section we review work on understanding user's mental models about security, and focus on two approaches that are amenable to modeling in software agents. We then describe representations of the models within a software agent that takes actions in its world that mimic those of a user who manages their own security, for example setting schedules for regular back-ups or virus scans. We show that the models can help predict regularities in behaviors observed in users.

A. Mental models extracted from the literature

Camp [9] finds that security experts predominantly use five kinds of mental models: physical, criminal, medical, warfare and market models. Non-expert users find physical and criminal models to be the most accessible [12]. Camp notes, however, that each model can evoke a different response from the user. Criminal models, for example, suggest investigation and prosecution by a central authority, while physical models emphasize lock-down and protection. Different models may therefore be appropriate for a user in different situations.

B. Mental models based on user survey

Wash [10] investigated the mental models that guided home computer users in deciding which expert security advice to follow. From structured interviews with 33 respondents, he identified eight models in two broad groups: 'viruses', a term used for any kind of malware, and 'hackers', used in any case where a human agent was envisaged in a potential attack. He found that around fifteen percent of the respondents had no particular model of a virus, while thirty percent view viruses as essentially buggy software that causes crashes (the *buggy* model), and a similar number views them as programs written by mischievous individuals impress their peers, causing harm to the infected computer (the *mischief* model). Another fifteen percent viewed viruses as programs written by criminals to gather sensitive financial information such as credit card numbers (the *crime* model). In contrast, every subject had some model of 'hackers'. Forty per cent saw them as opportunistic criminals looking for financial data, similarly to the criminal model of viruses (the *burglar* model). Roughly one quarter saw them instead as young, technically oriented and lacking moral restraint, breaking in to computers to cause damage and show off to their peers but not thieves (the *vandal* model). The remainding third also saw them as criminals, but targeting either rich or important individuals or large databases of information, and so not a threat to the respondent (the *big-fish* model).

We note that Wash found that individuals were able to give considerably more information about what hackers might do on accessing a host than they were able to give about a hacker's motivation or other characteristics. This provides anecdotal evidence for the position that the beliefs support mental models that are used to simulate the actions and resulting state when hackers or viruses may be involved in an attack. Wash went on to ask the subjects' opinion on the value of different kinds of security advice, such as using anti-virus software or regularly backing up files and compared the results with the mental models used. I will discuss these results further in the next section.

III. Implementation

Each of the mental models uncovered by researchers leads to patterns of behavior that allow a certain degree of prediction. We explored this by implementing mental models described by Wash in agent simulations. We follow Gentner and others [3] in viewing mental models as runnable, in the sense that they are simulated in the mind to answer questions about the world. In order to choose between alternative courses of action based on a set of mental models of their consequences for security, an agent simulates each alternative according to each of the models. Each combination results in a set of possible end states. The agent scores the end states using a fixed utility function and chooses the course of action that generally leads to the best score across the set of models.

Each mental model is represented as a set of operators that represent the features of the environment that will change when an action is taken. For example, if the agent were to install and use back-up software then, after this action is completed, the files in the agent's computer would be backed up and the agent would have less money, assuming the software was not free. The changes caused by an action are represented by a list of facts to be added to a state and a list to be deleted, in a style similar to STRIPS [13], although the changes can be conditional on existing state features. In addition to actions that the agent can take, the potential actions of third parties are represented in each mental model, along with 'trigger conditions', or logical statements about the environment that may cause the action to be taken. The actions may have probabilistic effects, producing a probability distribution of next states rather than a single next state. Agents can assess plans represented as sequences of actions against a mental model, or construct plans using search given an initial state and goal description.

A. Models of hackers and viruses

We show an example based on a question Wash asked of his subjects: whether it is advisable to make regular back-ups. We contrast the results obtained with the 'vandal' and 'burglar' mental models. For this question, the agent compares two plans that involve risky behavior, one that includes a step to back up data and one that does not. Each mental model is then used to simulate the possible outcomes of the plan, including potential actions of other parties and mitigating actions by the agent. Figure 1 shows a sketch of each of the simulated worlds for the plan that includes

backing up data under each mental model, as we describe below.

In the absence of mental models of hacker behavior, that state labeled S2 in Figure 1 will be the final state, and the inclusion of the action to back up data makes no significant difference except for the cost incurred to create the back-up. Based on the cost, our agents would probably reject the security action in this case, unless the back-up had intrinsic utility. Note that we do not specify the utility model here, since it is largely orthogonal to the mental models and the same agent choices will be made given a range of utility values.

When either the 'vandal' or 'burglar' mental model is used, more actions are posited as shown in the figure. In the 'vandal' model, the hacker may delete files, leading to state S3v. We use a simple turn-taking approach to simulate the evolution of the model, and so the agent next simulates steps it might take to restore the computer's state to a workable one. However this can only be done if a back-up was made before the attack, so in this model there is a clear advantage to the course of action that includes this step.

Next, suppose the agent compares these plans using the 'burglar' mental model. In this model there may still be an attack after the agent's initial plan is simulated, however the hacker does not delete files or crash the computer. Instead, he or she searches the computer for data that allows identity theft, leading to state S3b. Again the agent seeks actions to recover after the attack, but none are available whether or not a back-up was made, and the security action is not seen as valuable.

This example also serves to illustrate how mental models can capture correlations or independence between behaviors. Consider the additional security action of encrypting files on the computer hard drive. In contrast to backing up, this is likely to appear superfluous to a user with a 'vandal' model but useful to a user with a 'burglar' model. Therefore, if both individuals have one or other of these models, we would expect these behaviors to be negatively correlated, even though as complementary security actions they might be expected to be positively correlated. If mental models were distributed independently in the population we would expect no correlation. This observation reflects the findings of Aytes and Connolly [14], who queried 167 users about a number of security behaviors and found no significant correlation between the behaviors in general.

B. Validation

Eight mental models from Wash's work have been implemented and used to estimate the benefit of the security activities that he investigated. We compared the responses that agents would prefer using our models with those that Wash obtained from his study [4]. Specifically we considered whether four security activities were seen as worthwhile: using anti-virus software, exercising care in which website to visit, making regular backups and keeping patches up to date. These four were chosen because they lead to roughly even splits between those who thought they were important to follow and those who thought they could be ignored. In all cases we assume a utility model where the potential to avoid negative consequences will outweigh the cost of the security behavior if any consequences are predicted by the model.

The results are summarized in Figure 2. In this figure, a 'y' indicates that most users with the given model responded that the given behavior was important, and an 'n' indicates that most users responded that it was not important. The character is boxed if the mental model implementation leads to the same response. A blank cell indicates that there was no consensus response among users.

In our implementations, the 'crime' and 'burglar' models lead to prevention of access access but not to protection against vandalism, while the 'mischief' and 'vandal' models lead to protection, but not prevention of access. (The 'big-fish' and 'buggy' models do neither, though for different reasons.) This set of simple models does not predict that the 'mischief' and 'vandal' models would agree that it is important to show care in visiting websites. In general one would not expect a perfect match to observed user behavior, as human mental models are not necessarily self-consistent [15] and we have not modeled conditions under which the choice of model may be context-dependent.

IV. Discussion

We discussed mental models of security that have been elicited from users, and demonstrated an implementation that leads to decisions that match those of humans who report the same models. Mental models of security have been studied to improve communication or interface design, but to our knowledge this is the first time they have been used to model human security behavior. We believe that declarative implementations such as this will be useful ways for researchers to share the models they elicit and make them available to security researchers who want to test the impact of human behavior on security tools in development. Our approach uses a general-purpose model simulator in an agent platform that we plan to make available through the DETER project [16]. The use of a general-purpose simulator is intended to minimize the effort required for other groups to tailor our models or create new ones.

A. Related work

Several researchers have investigated mental models of security in addition to Camp and Wash described earlier. Dourish et al. [7] describe preliminary work with mental models with the aim of improving interfaces and communication. They found a dominant model of security as a barrier, a specialization of Camp's physical model. In addition to hackers, their subjects identified stalkers, spammers and

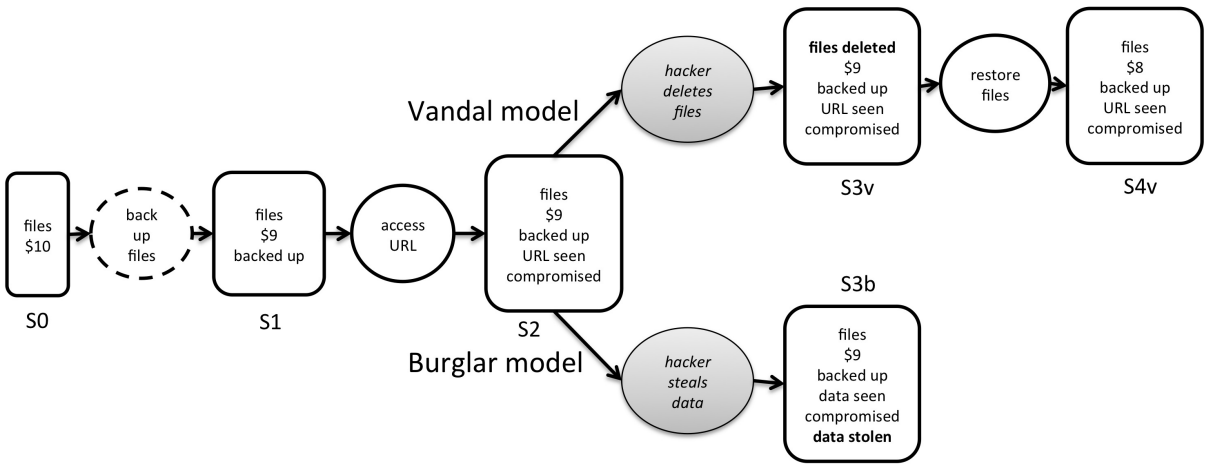

Figure 1. Simulations of mental models to decide whether to back up files, checked against the 'vandal' model of hackers (above) and the 'burglar' model (below). Each rectangle shows an initial or resulting state, with each line showing a different state variable. Each circle shows a possible action, where the darker circles are potential actions chosen by other actors. In the 'vandal' model, backing up allows files to be restored if they are deleted by the hacker and will probably be seen as worth the cost. In the 'burglar' model, it is seen as irrelevant since the hacker attempts to steal data does not delete files.

	buggy	mischief	crime	vandal	burglar	big-fish
use anti-virus-software	n		y		y	n
use care visiting websites	n	y		y	y	n
make regular packups		y	n	y	n	n
keep patches up to date		n	y		y	n

Figure 2. Behaviors predicted using the implemented mental models correlate with those given in responses from Wash's study.

marketers. Weirich and Sasse [8] investigate user beliefs in order to better persuade users to follow good practices. Aytes and Connolly take a rational agent and health communications perspective to structure knowledge about security [14]. Bravo-Lillo et al. elicited mental models as subjects responded to security warnings of different types, in order to create better warnings [17].

B. Future directions

In this paper we described mental models that directly cover the domain of cyber security. In many cases, however, individuals appeal to models that are not directly relevant but are more detailed and well tested, through a process of analogy. For example, explanations that appeal to medical or physical models of cyber security may very well lead individuals to use precisely these models to reason about security. Mechanisms for analogical reasoning with mental models have been studied extensively [18], and we plan to use them to explore how these analogies may lead to systematic patterns of decisions-making.

We also plan to explore how these models can improve adaptive interfaces that improve their communication with users about security over time. Since the models we use are declarative, they can provide a target for recognition, where the interface incrementally infers which models best describe the user and begin to tailor warnings and explanations based on the models. Through DETER, we plan to test the impact on networked attacks of large groups of individuals behaving according to a general population of mental models.

ACKNOWLEDGMENT

The authors are grateful for helpful discussions with colleagues from IU and the USC DETER and Game AI groups, including John Wroclawski, Steve Schwab, Jerry Lin and Marc Spraragen.

REFERENCES

[1] T. Benzel, J. Mirkovic, T. Faber, R. Braden, J. Wroclawski, and S. Schwab, "The deter project advancing the science of cyber security experimentation and test," in *IEEE HST Conference*, 2010.

[2] J. Blythe, A. Botello, J. Sutton, D. Mazzoco, J. Lin, M. Spraragen, and M. Zyda, "Testing cyber security with simulated humans," in *Innovative Applications of Artificial Intelligence*, 2011.

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

[4] R. Wash, "Folk models of home computer security," in *Proc Symposium on Usable Privacy and Security*, 2010.

[5] K. Craik, *The Nature of Explanation*. Cambridge University Press, 1943.

[6] P. Johnson-Laird, *Mental Models*, 1986.

[7] P. Dourish, J. Delgado De La Flor, and M. Joseph, "Security as a practical problem: Some preliminary observations of everyday mental models," in *CHI Workshop on HCI and Security Systems*, 2003.

[8] D. Weirich and M. A. Sasse, "Pretty good persuasion: a first step towards effective password security in the real world," in *2001 Workshop on New security paradigms*. ACM, 2001.

[9] L. Camp, "Mental models of privacy and security," *Ieee Technology And Society Magazine*, vol. 28, no. 3, 2009.

[10] R. Wash and E. Rader, "Influencing mental models of security: A research agenda," in *New Directions in Security and Privacy*, 2011.

[11] J. Blythe, J. Camp, and V. Garg, "Targeted risk communication for computer security," in *Proc. International Conference on Intelligent User Interfaces*, 2011.

[12] V. Garg and J. Camp, "How Safe is Safe Enough: Online Version," in *Security and Human Behavior*, 2010.

[13] R. E. Fikes and N. J. Nilsson, "Strips: A new approach to the application of theorem proving to problem solving," *Artificial Intelligence*, vol. 2, no. 3, pp. 189 – 208, 1971.

[14] K. Aytes and T. Connolly, "Computer security and risky computing practices: a rational choice perspective," *Journal of Organizational and End User Computing*, 2004.

[15] D. Norman, *Some Observations on Mental Models*, 1983.

[16] "The deter project," http://deter-project.org.

[17] C. Bravo-Lillo, L. Cranor, J. Downs, and S. Komanduri, "Bridging the gap in computer security warnings: A mental model approach," *Security Privacy, IEEE*, vol. 9, no. 2, 2011.

[18] B. Falkenhainer, K. D. Forbus, and D. Gentner, "The structure-mapping engine: Algorithm and examples," *Artificial Intelligence*, vol. 41, no. 1, pp. 1 – 63, 1989.