Modeling and Integrating Cognitive Agents Within the Emerging Cyber Domain

Randolph M. Jones[1], Ryan O'Grady[1], Denise Nicholson[1], Robert Hoffman,[2]
Larry Bunch[2], Jeffrey Bradshaw[2], and Ami Bolton[3]

[1]Soar Technology
Ann Arbor, MI

[2]IHMC
Pensacola, FL

[3]Office of Naval Research
Arlington, VA

rjones@soartech.com, ryan.ogrady@soartech.com, denise.nicholson@soartech.com,
rhoffman@ihmc.us, lbunch@ihmc.us, jbradshaw@ihmc.us, amy.bolton@navy.mil

ABSTRACT

One of the elements missing from virtual environments in the emerging cyber domain is an element of active opposition. For example, in a training simulation the instructor assigns the student a task or objective, and the student then practices within the environment (the “cyber range”) until they feel comfortable with the task or are able to demonstrate the requisite level of mastery. The environment may have static defenses, such as access control or firewalls, or a fixed set of intrusion methods to defend against, but it typically lacks any active opposition that might adapt defensive or offensive actions (e.g., monitor logs, blocked connections, exploit switching or information gathering). This is akin to training fighter pilots against adversaries who know how to use their weapons, but do not have any tactical or strategic goals beyond that. This is unfortunate for two reasons: 1) it trains cyber operators to behave as though opponents do not have a tangible existence or do not have higher-level goals, and 2) it ignores an opportunity to tailor the student’s learning experience through adjustable adversary behavior. Cognitive agents have the potential to transform the cyber operations training experience. The application of cognitive agents to the roles of cyber offense and defense would provide a more complete cyber ecology for training purposes and thus a more realistic training experience for the student. There are two key challenges to creating such cyber agents: 1) modeling the complex, and continually evolving, processes of cyber operations within a cognitive architecture, and 2) defining the tools and data standards to enable cognitive agents to interoperate with networks in a portable way. This paper discusses novel models of cyber offensive and defensive behavior based on observation and elaboration of human expertise, as well as an approach to the creation of software adapters that translate from task-level actions to network-level events to support agent-network interoperability.

ABOUT THE AUTHORS

Randolph M. Jones, PhD, is a senior artificial intelligence engineer at Soar Technology, and co-founded Soar Technology in 1998. Dr. Jones received his BS in mathematics and computer science from UCLA, and he received his M.S. and Ph.D. in information and computer science from the University of California, Irvine.

Ryan O’Grady is the technical lead for Soar Technology’s emerging business area in cyberspace training and visualization, and a senior software engineer in the Intelligent Training business area. Mr. O’Grady received a BSE in Computer Science Engineering from the University of Michigan in 2004. Certifications: Security+, CPTE, OSCP

Denise Nicholson, PhD, CMSP, is the Director of Soar Technology’s new Technology Area "X" leading an effort to explore, identify and pursue innovative applications of intelligent systems for critical and challenging problems, such as Cyber Security. Dr. Nicholson has a Ph.D. and M.S. in Optical Sciences from the University of Arizona, and a B.S. in Electrical Computer Engineering from Clarkson University.

Robert Hoffman, PhD, is a Senior Research Scientist at the Florida Institute for Human and Machine Cognition (IHMC). He is senior editor of the Department on Human-Centered Computing of *IEEE: Intelligent Systems*. His latest book is Accelerated Expertise: Training to High Proficiency in A Complex World (2014, Taylor & Francis).

Larry Bunch is a Senior Research Associate at IHMC. He received his BS in computer science from the University of West Florida and has published extensively concerning software agents, semantic policies and reasoning, and large-scale event visualizations.

Jeffrey M. Bradshaw, PhD, is a Senior Research Scientist at IHMC. He co-edits the HCC Department of IEEE Intelligent Systems and has published widely in software agents, semantic technologies, digital policy management, and human-agent-robot teamwork (HART).

Ami Bolton, PhD, is a Program Officer at the Office of Naval Research (ONR). Her programs focus on enhancing individual and team decision-making and combat effectiveness through advances that improve perception, cognition, and team coordination. Dr. Bolton received a M.S. in Human Factors from the Florida Institute of Technology, and Ph.D. in Applied Experimental & Human Factors Psychology from University of Central Florida.

Modeling and Integrating Cognitive Agents Within the Emerging Cyber Domain

Randolph M. Jones[1], Ryan O'Grady[1], Denise Nicholson[1], Robert Hoffman,[2]
Larry Bunch[2], Jeffrey Bradshaw[2], and Ami Bolton[3]

[1]Soar Technology
Ann Arbor, MI

[2]IHMC
Pensacola, FL

[3]Office of Naval Research
Arlington, VA

rjones@soartech.com, ryan.ogrady@soartech.com, denise.nicholson@soartech.com,
rhoffman@ihmc.us, lbunch@ihmc.us, jbradshaw@ihmc.us, amy.bolton@navy.mil

Cyber warfare presents a persistent and evolving threat to military and civilian information systems. Both DoD (Parrish, 2013) and ODNI (Pellerin, 2013) rank cyber warfare as our top national security concern. In addition to threats to our defensive forces, cyber attacks pose an economic threat on the order of one trillion dollars (Ponemon, 2013). Although individual cyber-warfare tools operate at extremely fast speeds, aggressors increasingly pursue a "cyber kill-chain" (Hutchins et al., 2010) over days, weeks, or months. Would-be cyber aggressors are constantly changing their attack vectors to take advantage of security lapses by human resources and the latest vulnerabilities in information technology. These human-speed activities are guided by cognitive behavior that includes a variety of types of goals and expertise: script-kiddies, ideological activists, investigators, financial criminals, intelligence agents, or cyber warfighters (Lathrop et al., 2010). At the human, cognitive level, offense depends on and reacts to responses of defenders (Pfleeger & Caputo, 2012) and users (Bowen et al., 2012) that are also cognitively driven. Current cyber-warfare tools comprise suites of technical mechanisms that respond to the *tools* that aggressors and defenders use, but *not to the individuals themselves*. Human tactics are currently addressed through human-staffed wargames at cyber ranges (Merit, 2013; Pridmore, 2012). Human role-players are expensive, not repeatable, and not deployable as an automated system. There is an emerging need for cognitive-level synthetic cyber offense and defense, to ensure realistic cyber simulation and training.

Building effective training systems for cyber warfare presents a suite of unique problems:

- Offensive and defensive activity is highly dynamic.
- The characteristics of target network environments are driven by the users of the system and their current activities, which are highly variable and unpredictable.
- User behavior often creates vulnerabilities that can be exploited.
- Cyber warfighters themselves are extremely adaptive and creative. In order to meet their objectives they will change tactics or tools based on opportunities detected in a computer network or responses initiated by adversaries or users.

Current training environments do not adequately capture the dynamic and cognitive-level characteristics of cyber warfare. They are unable to capture the purposefulness, creativity, and adaptability of actual cyber warfighters. Studying previous offensive and defensive scenarios in a classroom environment is an effective means of understanding the building blocks of cyber warfare, but falls short of creating the skills needed to deal with a creative and time-sensitive event or a sophisticated but dynamic plan. Computerized unit tests can build fundamentals for dealing with individual components cyber warfare, but they do not help the trainee learn to recognize and make sense of the larger picture, nor do they capture the dynamic nature of networks and users.

If cyber warfighters are to learn to respond to a cunning and adaptive opponent, they need to train against cunning and adaptive opponents. An effective cyber-warfare training system must be adaptable and deal with the changing nature of a networked environment. It must be able to model the dynamic nature of cyber aggressors, users, and defenders. It must create a virtual environment that replicates the environment that the trainee will ultimately operate in. An appropriate virtual environment also creates the opportunity for accurate post-event forensic analysis by providing access to databases, configurations, and system logs. This paper presents our efforts to address these issues through the development of cognitive agents for cyber offense and defense.

The Soar cognitive architecture described in this paper is not to be confused with Soar Technology, the affiliation of some of the authors. Soar is not a commercial product, but is available under a General Public License from http://soar.eecs.umich.edu/ maintained by the University of Michigan. The Soar architecture provides the technological foundation for the cognitive agents described here.

MODELING AND INTEGRATION CHALLENGES

In order to build realistic cognitive agents, the agents must encode appropriate domain expertise, and they must interact with a realistic cyber environment (Jones & Laird, 1997). In addition to realism, cost effective cognitive agents also need to address these related issues:

- Reduce cost of realistic role playing in cyber-warfare simulation, system engineering, and analysis of cyber operations.
- Enable end-user updating of agent knowledge with minimal support from software engineers, both through coaching by instructor Subject-Matter Experts (SMEs) and through explicit addition of new knowledge about cyber tactics.
- Be readily adaptable to a wide range of network structures, devices, and protocols.

In order to achieve these goals, we have built cyber cognitive agents from a prior infrastructure for cognitive modeling of other types of tactical decision making (e.g., Jones et al., 1999). In this section, we describe the challenges and our approach to the problems of modeling and integration.

Cognitive Modeling

Human cognitive behavior is rich, subtle, and combines autonomous and deliberative information processing. The challenge in developing an executable cognitive agent is formalizing information from human experts and documentation, and embedding it in a computational framework that is sufficiently rich to manipulate beliefs, goals, and actions in ways that correspond to humans.

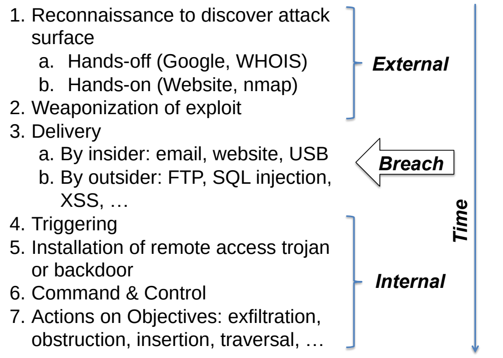

1. Reconnaissance to discover attack surface
a. Hands-off (Google, WHOIS)
b. Hands-on (Website, nmap)
2. Weaponization of exploit
3. Delivery
a. By insider: email, website, USB
b. By outsider: FTP, SQL injection, XSS, …
4. Triggering
5. Installation of remote access trojan or backdoor
6. Command & Control
7. Actions on Objectives: exfiltration, obstruction, insertion, traversal, …
External
Breach
Internal
Time

Figure 1: Cyber Kill Chain: Initial efforts external to the target discover the attack surface, which is then breached to allow internal actions.

One key thrust of cognitive agent design is *cognitive task analysis* (CTA; Hoffman & Militello, 2008). CTA is an empirical and rational analysis of the information flow, constraints, and task structures, which dictate the sets of plausible decisions and actions that will achieve relevant goals. To fill a fully realistic cognitive environment, we must analyze tasks for users (to generate realistic background traffic), aggressors, and defenders. For these projects, we develop the CTA within the principles of the Soar cognitive architecture (described below). With CTA providing the task framework, we then represent knowledge and expertise, acquired either from SMEs, publications produced by SMEs, or other case studies and documentation. The addition of knowledge from SMEs allows us to extend *plausible* cognitive agents (constrained by the CTA) into *realistic* cognitive agents (reflecting actual human expertise). Using methods for knowledge acquisition (another form of CTA), we have analyzed bodies of knowledge relevant to offensive and defensive cyber activities. Example source materials include published analyzers, including an analysis of the cyber "kill chain" (Hoffman 2010; see Figure 1) and publications of the SANS Institute (Forsythe et al., 2012). In addition, the Michigan Cyber Range (MCR) and the Human Effectiveness Directorate at AFRL serve as technical advisors to our research.

Ultimately, cognitive agents express task knowledge in the form of a series of actions that are conditioned on combinations of situational beliefs and goals. This requires developing a situation-representation language with the support of SMEs. To achieve this, the model builders and SMEs work together on increasingly complex scenarios, ultimately reformulating the SME's informal verbalizations of knowledge in the formal, structured representation that the cognitive agent executes. Our cognitive modeling process relies on a variety of question types that are iterated over each decision point in a scenario to produce a sophisticated knowledge base that is adaptive to a wide variety of goals and situations.

Our approach to development recognizes commonalities and differences between offensive and defensive agents. There is a body of situation-understanding knowledge that is shared by both types of agents. Specific tactical knowledge tends to be different across agents, although there are some types of tactics that can see use both offensively and defensively, especially when we consider the maxim that "the best defense is sometimes a good offense". We exploit opportunities to reuse knowledge across agents whenever possible, resulting in cost-effective development, as well as robustly engineered and modular adaptive systems.

Exploiting Cognitive Architecture

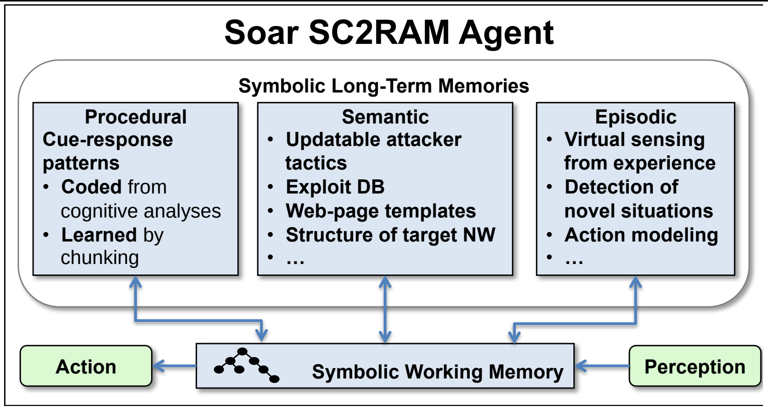


Figure 2: Cyber Agent in the Soar cognitive architecture.

An additional tool for building realistic cognitive agents is to build from an established cognitive architecture. Cognitive architectures are theories of mind instantiated as software programming environments, from which we can build executable cognitive agents that interact with real, virtual, or simulated cyber networks in real time. There are alternative cognitive architectures to choose from, but we have selected the Soar architecture for this work, because it provides us with several advantages. Soar incorporates decades of experience in cognitive research into an architecture that is reusable across new cognitive agents. Soar agents integrate varieties of knowledge, learning, and reasoning strategies, including semantic and episodic as well as procedural memory, reinforcement learning, spatial reasoning, and activation dynamics, as well as cognitive agents that depend on situation understanding and interaction with complex environments (Laird, 2012). In addition to providing a reusable software architecture for the development of cognitive agents, Soar is grounded in a unified theory of cognition. In spite of its cognitive realism, the basic Soar engine is compact (the Java binaries are only 11 MB, and there is a version that runs on an iPhone), and fast (executing basic decision cycles an order of magnitude faster than the 50 ms required by the human brain, Jones et al., 2011).

Soar provides a fixed perception-action cycle and several memory models, which specific cognitive agents instantiate with task- and domain-specific knowledge. The architecture itself does not commit to any particular reasoning strategies, so the job of the cognitive modeler is to identify the representations, goals, and strategies used by the human experts that are the target of the model. Figure 2 illustrates a portion of this representation for cyber agents. The basic Perceive-Decide-Act cycle includes Soar’s Symbolic Working Memory (where active reasoning takes place), external Actions that are invoked in the environment, the network environment in which the agents are embedded, and network cues that are monitored on the input link by the Perception module.

The figure also illustrates three specialized long-term memories that encapsulate different types of knowledge. *Procedural memory* stores cue-response patterns. Many of these are derived from our cognitive task analysis, but Soar can derive others through its chunking process. When a Soar agent encounters a situation for which it has no pre-defined patterns, it generates a sub-state in which it reasons more deeply about the situation, including exploring alternative futures to evaluate possible next steps. It stores a trace of this reasoning, and can convert the conditions that spawned it, and the outcome that it finally selected, into a new pattern. The next time it encounters a similar situation, this pattern is immediately available. Once the initial agent knowledge is in place, chunking allows the agent to extend that knowledge with new patterns that reflect its experience in problem solving.

Semantic knowledge is declarative knowledge that can often be compiled from tables or human interfaces manipulated directly by domain experts (rather than programmers). This feature makes semantic memory the key to keeping the agents maintainable and up to date. Knowledge structures can be parameterized declaratively, allowing updating of the agent by instructional staff rather than programmers. Semantic memory contains a database of exploits (for offense), tasks (for users), and defensive actions (for defense). The agents map perceptions and situation interpretations into these databases to identify appropriate actions.

Episodic memory remembers the agent’s recent experiences, enabling it to recognize novel situations (those that do not match remembered states), apply “virtual sensing” against past situations similar to the current one, and anticipate the result of actions under consideration. We use episodic memory in the cyber agents to identify novel responses to agent actions, and then attempt to integrate those new responses into innovative exploits against the adversaries.

Simulation Realism and Integration

The final thrust for building realistic cognitive agents is to situate them into a realistic environment. This requires a focus not only on the simulation environment itself, but an agent middleware that provides a reusable level of abstraction to allow interoperation with multiple environments. To do this we rely on IHMC's Sol cyber framework for a variety of tasks (Bradshaw et al., 2012). For instance, Sol's innovative Network Observatory visualization is used to make the unrolling of complex events comprehensible (Bunch et al., 2015). In the first phase of the project, we relied on a simulated network and synthesized cyber-attack events, but now have migrated to Merit's Michigan Cyber Range (Merit, 2013). Because of the abstractions provided by Sol's tools and ontologies, replacing the simulated network with the cyber range was straightforward. By embedding the agents in multiple, realistic environments, we have been able to develop an agent middleware that supports the use of synthetic cyber tools (software simulations) as well as actual cyber toolkits, such as NMAP. The use of actual cyber toolkits acts as a forcing function to ground the realism of the cognitive agents from the bottom up, with the CTA and subject-matter expertise enforcing realism from the top down. Our integration results thus far show that the Michigan Cyber Range is suitable for a wide range of cyber systems simulation. In addition, our success in integrating with a real cyber range smooths the transition path to other existing cyber ranges for testing, evaluation, and training.

COGNITIVE AGENT IMPLEMENTATION

Having described the top-down and bottom-up constraints and techniques for developing realistic cognitive agents, we now delve into some of the details of the agents we have developed. Space limitations do not allow us to provide a high level of detail on both the offensive and defensive agents, but they share modeling approach, knowledge representation, and some specific knowledge elements.

As described above, the cognitive agents implement knowledge representations and reasoning strategies we have elicited from human subject-matter experts and existing documentation and case studies, tools, and methods. The agents' generative and adaptive abilities are derived from a combination of ontological knowledge of cyber infrastructure and tools, together with domain-specific task knowledge that is linked by cyber-warfare goals and constrained by the ontological knowledge. Domain-specific knowledge is supplemented with domain-independent knowledge (e.g., for deception, tactical analysis, and various forms of learning), allowing the agents to synthesize strategies at run time in response to the dynamics of the current situation, just as human cyber-warfighters do. The agents use a least-commitment reasoning strategy that allows them to stick with long-term goals and plans as long as they are relevant, but to adapt goals, plans, and individual tactics as the agents gather new information about infrastructure changes and adversary responses.

Initial agent development has served two purposes: to provide a proof of concept of the modeling approach, and to provide the architecture of emerging agent extensions. The initial agent implementations include two primary types of knowledge. Procedural knowledge includes a goal/task hierarchy derived from cognitive task analyses and subject-matter expertise, while declarative knowledge accumulates and draws on knowledge of the network configuration acquired during operation.

The goal/task hierarchy of the initial agents includes multiple levels of knowledge, such as high-level goals (e.g., disrupt-operations), mid-level goals (e.g., disrupt-service-source, achieve-presence-on-machine), and primitive actions (e.g., vertical-scan, denial-of-service-attack, send-payload) that the agents can execute via scripts or actual tools provided by a cyber-toolbox. The highest-level goals are assigned by the human supervisor who deploys the agents, and different agent instances on a team can have different high-level goals. Goals and actions are associated in memory with their potential effects. This association allows flexible, situation-dependent composition of plans to achieve goals (generative creativity). The agents will not adopt a goal if it does not have actions that can achieve the goal. The association also produces strategy switches in response to failures, successes, or new observations, and allows the agents to learn new "potential effects" through experience by monitoring actual effects of actions (learned creativity).

Figure 3 shows the goal hierarchy implemented in the initial model of cyber offense. Ellipses are high- and mid-level goals, while rectangles are primitive actions. The red rounded rectangles illustrate how these actions map to ploys for deception that will be used to extend the agent's generative capabilities. The agent repeatedly evaluates its currently active goals in the light of the state of the network, selects subgoals that could achieve its desired effect, and activates them, thus taking action in the network. Thus the sequence of actions is not hard-coded into the agent, but emerges based on goal-means reasoning that captures the expertise of human attackers. The human-like nature of this reasoning has two important benefits. 1) The agent can explain its actions in a way that people can understand, a valuable capability when using the agent to audit a defensive automation system. 2) The agent's reasoning can be incrementally upgraded by a human supervisor when the agent finds itself at a dead-end.

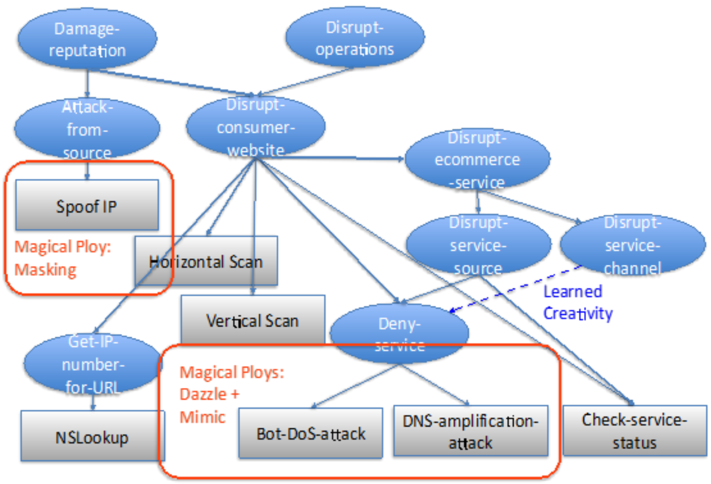


Figure 3: Goal Hierarchy in an initial cyber-offense cognitive model

Figure 4 illustrated the initial goal decomposition for a model of cyber defense. The two agents share the code they use for sensemaking and goal management. The slightly different "shapes" of the goal hierarchies reflect, in part, different points of emphasis in our development of the two agents, but also some difference in the nature of offensive and defensive missions. For example, cyber defense often centers on monitoring the baseline behavior of the network and trying to diagnose discrepancies to determine whether they are benevolent or hostile. Cyber offense can include monitoring and diagnosis activities as well, but they do not always play as central a role.

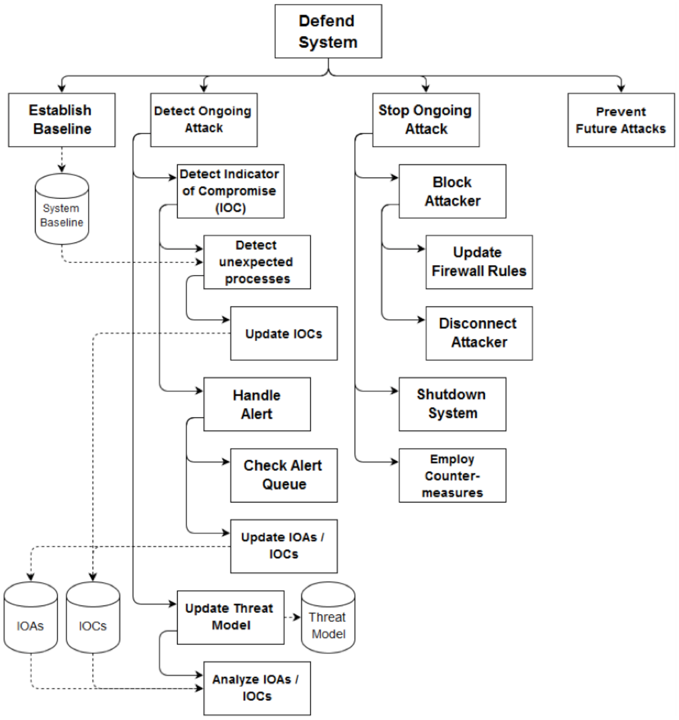


Figure 4: High-level defender goal hierarchy

The two most significant functional types of knowledge in the agents are *task knowledge* and *situational knowledge*. The basic units of *task knowledge* in the agents are goals, subgoals, and tactical methods. These knowledge units are tagged with (sometimes numerous) associative links that indicate the situations in which the knowledge is relevant, the potential super-goals that the knowledge can be used to achieve, potential adversary responses, and other types of indices. *Situational knowledge* allows the agents to interpret observable information (in the context of active goals and expectations) to paint a continuously updated picture of situational understanding. Associative links between situation understanding and task knowledge allow the agents to use a least-commitment strategy to adapt tactics as adversary responses unfold. Associative links to a library of deception strategies allow the agents to create and achieve deception goals in parallel with the primary tactical goals. When these basic knowledge units interact with large ontological databases describing cyber infrastructure and available offensive and defensive tools, the agents will be able to generate a wide variety of adaptive and realistic threads of offense or defense. With the initial reasoning architecture already in place, future work will involve integrating the architecture with new ontologies, enriching the task and situational-understanding knowledge from subject-matter resources, and integrating new, sophisticated deception and learning strategies. We are extending previous work to develop suitable OWL ontologies (e.g., MIT/LL and IHMC, 2013). This growing set of ontologies, coupled with cyber tools our team has also developed, insulates cognitive agents from low-level descriptions of network events and the details of agent actions to be performed. We accomplish this by interpreting and elevating the semantics of events and actions to an appropriate "human-readable" level.

COGNITIVE ADAPTIVITY AND LEARNING

Our initial agents incorporate fairly simple logic for choosing offensive and defensive actions, but they do not yet include a comprehensive understanding of how to compromise or defend computer systems. The initial agents provide a conceptual structure in which to organize this knowledge, based on a cognitive analysis of deception, sensemaking, and relevant features of uncertainty and complexity. The initial agents also implement an extensible, first-principles knowledge representation. We are building from this base representation and additional analyses to implement agent extensions that naturally synthesize complex offensive and defensive strategies.

Our current work focuses on building a more thorough body of cyber-relevant knowledge, and developing a grammatical model of cyber warfare to facilitate the continual updating of agent knowledge. Our analysis of cyber operations suggests that we can capture much of the structure of cyber strategy as a formal model. The key focus is to develop an ontological model that relates abstractions about tactical and deceptive approaches and constraints that govern how they may be combined, to kinds and permissible sequences of specific tactics. This will allow the cognitive agents to generate novel attacks from first principles, reasoning as appropriate about possible moves at multiple levels of abstraction.

This approach will also form the foundation for learning mechanisms that minimize the need for model engineers to update the agents manually through time. Our initial work has focused on two kinds of learning:

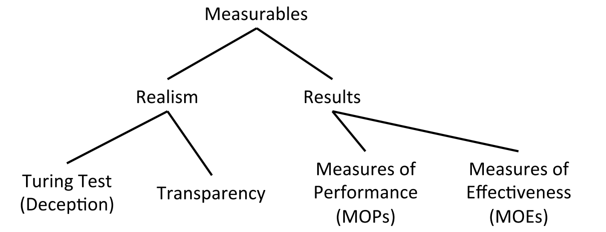
- *Inferential learning* is the result of Soar's goal-driven reasoning process, and occurs when this process derives a new high-level insight from previously known low-level steps. In the Soar architecture, this capability is called "chunking." It is analogous to the learning that a student experiences on working through a proof: the student "knows" all the pieces, but inferential learning leads her to recognize their implications. In our initial agents, inferential learning takes place when the agents dynamically compile multiple steps to generate a coherent attack.
- *Observational learning* results from agent interaction with the external world. For example, our initial offensive agent deploys a distributed denial of service attack for the goal of disrupting a service *source*, and learns from experience that the same action can also accomplish the goal of disrupting a service *channel* when the defensive response includes blocking traffic from the attack addresses. Architecturally, Soar supports this kind of reasoning through a combination of its semantic and episodic memories.

We are further exploring two additional forms of learning:

- *Abductive learning* is reasoning from observations to the best explanation. This mechanism will allow the agents to infer causal explanations to observed adversary responses, and then use these causal models to innovate new attacks.
- The development of a formal knowledge representation also opens the door to *instructional learning*, in which a human supervisor can coach individual instances of the agents. We investigate this form of learning in the context of prior work on taskability and interactive task learning (Huffman & Laird, 1993; van Lent, 2000).

In all four types of learning, as an agent instance learns, its can share its knowledge with other instances, either on the same problem or across multiple scenarios, so that the agent knowledge bases increase in capability over time with minimal need for support by model builders.

EVALUATION


Measurables
Realism
Results
Turing Test (Deception)
Transparency
Measures of Performance (MOPs)
Measures of Effectiveness (MOEs)
Figure 5: Evaluation measures

As our cyber agents mature, a major future activity will be the development of operational definitions and measurement scales. Figure 5 outlines a taxonomy for our evaluation approach, with our initial emphasis on Transparency and MOEs. The fundamental objective for our agents is to provide cognitively realistic models of human cyber warriors that can reduce the cost of human role playing in training environments. Evaluation can focus either on the cognitive *realism* of the agent, or on the *results* it produces.

Measures of *realism* deal primarily with how human-like the agent behavior is. The Soar community has developed methods that can assess this aspect of realism (Laird et al., 2001), asking human judges to evaluate whether a behavioral trajectory is being driven by a human or a computer. However, recent research (Gratch & Marsella, 2004) suggests that such tests are weak, because agents designed to be unrealistic can actually score higher than those that are designed to be realistic. In addition, many aspects of human realism, such as decision biases and errors, would generate a more realistic agent but not one that is more effective as a member of a training role-playing team. From the point of view of DoD operational utility, the ability of the agent to conceal its computational nature is relatively unimportant. An alternative measure of realism is more useful in operation, and that is the agent's *transparency*; the ability of humans to understand its reasoning. This feature is particularly important in training systems, where a human supervisor needs to be able to interact with the agent to assess the current state of its tactics and suggest modifications.

Evaluation of a system's *results* is commonly discussed in terms of Measures of Performance (MOPs), which can be defined within the boundaries of the system under test, and Measures of Effectiveness (MOEs), which require embedding the system in an environment (Sweet et al., 1985). In the case of our cyber agents, MOPs (such as time to reach a decision) are less important than MOEs. The reason is that one great benefit of a cognitive agent is its ability to adapt and innovate in the back-and-forth of cyber operations, a capability that requires embedding in the operational environment and thus is a MOE. In our research to date, we have identified a number of potentially useful MOEs for evaluating the initial cyber agent implementations and for guiding future development.

An additional approach to evaluation is to estimate the "face validity" of our agents on a variety of scenarios. We have previously investigated the utility of cognitive evaluations that combine "face validity" with realistic simulation environments and scenarios. A more fine-grained approach to evaluation captures transcripts of expert role players, and records their decisions in protocols. We have explored methods for mapping an agent's individual goals, decisions, errors, and strategy shifts to data collected from human individuals (e.g., Jones et al., 1999). We will investigate the best of these approaches for evaluation of the agents.

SUMMARY OF COGNITIVE MODELING STRENGTHS FOR CYBER WARFARE

In situating our cognitive agents for cyber warfare in the extensive current research landscape on cyber operations, it is useful to make several distinctions. Our agents are cognitive (vs. non-cognitive), generative (vs. diagnostic), based on a mature architecture for cognition (vs. ad-hoc), and on-line (vs. off-line). First, the agents are explicitly *cognitive* models of cyber operations. Though useful for other purposes, most cyber-operations tools and approaches do not leverage an understanding of human cognition in significant ways. One example is learning and matching signatures of specific exploits, as in the DARPA Cyber-Genome or Cyber Grand Challenge programs, which rely largely on non-cognitive machine-learning and pattern-matching technologies. Another is the automation of specific offensive and defensive functions, such as port scanning, password cracking, or traffic filtering. Such functions can benefit from technical advances such as parallel processing architectures, and the toolkits available to our agents can take advantage of these capabilities. However, the agents themselves implement the human reasoning that lies behind the deployment of these capabilities, allowing them to employ strategies that take advantage of context and adversarial reasoning techniques. A cognitive approach is important in dealing with advanced persistent threats because such threats consist of multiple steps, chosen over time by the adversary, depending on the adversarial objectives and the current state of target systems that are being defended. Such goal-based, context-sensitive decision-making is fundamentally cognitive in nature.

Second, the agents are *generative* (as opposed to diagnostic) models. Diagnostic models seek to marshal relevant evidence and help the human to understand the state of the battlespace. The extensive body of work on cyber situation-assessment (e.g., Liu et al., 2010; Mahoney et al., 2010) is fundamentally diagnostic in nature, using static representations such as fuzzy cognitive maps to capture the relations among key observables. The fundamental challenge in constructing such a diagnostic system is the creativity of the cyber warfighter. Any diagnostic representation of behavior quickly becomes obsolete as cyber warfighters modify their actions in response to evidence that they are being detected. Generative mechanisms can create new, previously unseen tactics in three ways. First, they can combine different low-level offensive or defensive elements in previously unseen combinations (constrained in ways that reflect domain knowledge) as needed to address high-level goals. Second, the cognitive agents can learn from experience, recognizing unanticipated effects of actions and exploiting them in later efforts. Third, because the agents are cognitively realistic, they can explain their reasoning to human supervisors and readily accommodate new suggested exploits or defenses, continuously expanding their knowledge by interaction with users. The ability to generate an unlimited space of offensive and defensive tactics provides a necessary prerequisite to crafting diagnostic mechanisms for cyber warfare.

Third, among alternative generative approaches to cognitive cyber automation (e.g., Pokorny, 2010; Weyhrauch, 2013), the agents are based on the *mature* Soar cognitive architecture, which over several decades has proven its ability to provide realistic human behavior in applications as complex and challenging as piloting a fighter aircraft (Jones et al., 1999) or populating a virtual world with culturally realistic characters (Sims & Taylor, 2009). It is popular to characterize an algorithm as “cognitive” based on its representational structure, e.g., whether it has explicit structures for beliefs, desires, and intentions, but these structures may not be readily understandable by human partners. Soar not only has cognitively realistic representations, but also has demonstrated its ability to interact with humans in complex dynamic environments, as a result of decades of application and refinement in a wide range of applications.

Fourth, the agents are designed as *online* tools, intended to be applied to real systems in real time. This commitment sets the cognitive agents apart in two ways. First, the agent interaction middleware allows the agents to be connected to real or simulated networks to address the pressing need for more efficient integration and deployment. The rich semantics of the models of “live” networks enables the agents to reason effectively and powerfully about strategies, moves, and effects. Extensions to the middleware and ontologies, including those created to support the MIT Lincoln Lab CCER Interoperability Standards (MIT/LL and IHMC, 2013), among others, will enable interoperability with national cyber ranges and will facilitate adoption by government transition partners. A second characteristic, distinguishing our agents from other approaches, is that they are designed to run non-stop, real-time in a production environment rather than as a stand-alone experimental framework for testing cognitive theories.

This combination of distinctions, together with our development approach based on cognition and integration, provides a strong platform from which to launch further research into training applications for cyber warfare. As we have argued, the cognitive element is essential to future realistic training in the emerging cyber domain, and realistic cognitive agents will play a key role in this training.

ACKNOWLEDGEMENTS

The work presented here has been supported in part by ONR, Simulated Cognitive Cyber Red-team Attacker Model Phase II, contract # N00014-15-C-0100. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defense or Office of Naval Research. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here. This work benefited from contributions by Van Parunak.

REFERENCES

B.M. Bowen, S.J. Stolfo, et al. Measuring the Human Factor of Cyber Security. *Homeland Security Affairs, IEEE 2011 Conference on Technology for Homeland Security: Best Papers*, Supplement 5, 2012.

J. M. Bradshaw, M. Carvalho, L. Bunch, T. Eskridge, P. Feltovich, M. Johnson, and D. Kidwell. Sol: An Agent-Based Framework for Cyber Situation Awareness. *Künstliche Intelligenz*, 26(2):127-140, 2012. http://www.jeffreymbradshaw.net/publications/SOL.pdf.

L. Bunch, J.M. Bradshaw, R. R. Hoffman, and M. Johnson. Principles for Human-Centered Interaction Design, Part 2: Can Humans and Machines Think Together?. *IEEE Intelligent Systems*, 30:3, 68-75.

J.C. Forsythe, A. Silva, S. M. Stevens-Adams, and J. M. Bradshaw. Human Dimensions in Cyber Operations Research and Development Priorities. Sandia National Laboratory, Albuquerque, NM, 2012.

J. Gratch, S. Marsella. A Domain-independent Framework for Modeling Emotion. *Journal of Cognitive Systems Research*, 5(4):269-306, 2004.

R.R. Hoffman. Theory → Concepts → Measures but Policies → Metrics. In E. Patterson and J. Miller (eds.),, *Macrocognition metrics and scenarios*. London: Ashgate, 3-10, 2010.

R. Hoffman & L. G. Militello. Perspectives on Cognitive Task Analysis: Historical Origins and Modern Communities of Practice. Boca Raton, FL: CRC Press/Taylor and Francis, 2008.

S.B. Huffman, J.E. Laird. Instructo-Soar: Learning from Interactive Natural Language Instructions (Video Abstract). In *Proc. AAAI*, pages 857, 1993.

E.M. Hutchins, M.J. Clopperty, et al. Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains. Lockheed Martin Corporation, 2010.

S. Jajodia, P. Liu, et al., Editors. *Cyber Situation Awareness: Issues and Research*. Advances in Information Security, New York, NY, Springer, 2010.

R.M. Jones, J.E. Laird. Constraints on the design of a high-level model of cognition. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 358-363). Stanford: Psychology Press, 1997.

R.M. Jones, J.E. Laird, P.E. Nielsen, K.J. Coulter, P. Kenny, F.V. Koss. Automated intelligent pilots for combat flight simulation. *AI Magazine*, *20*(1), 27–41, 1999.

R.M. Jones, S. Furtwangler, M. van Lent. Characterizing the performance of applied intelligent agents in Soar. *Proceedings of the 2011 Conference on Behavior Representation In Modeling and Simulation (BRIMS).* Sundance, UT, 2011.

J.E. Laird. *The Soar Cognitive Architecture*. Cambridge, MA, MIT Press, 2012.

J.E. Laird, J.C. Duchi. Creating Human-like Synthetic Characters with Multiple Skill Levels: A Case Study using the Soar Quakebot. In *Proc. AAAI Spring Symposium*, AAAI, 2001.

S. Lathrop, J.M.D. Hill, et al. Modeling Network Attacks. In *Proc. 12th Conference On Behavior Representation In Modeling And Simulation (BRIMS 2003)*, 2003.

S. Mahoney, E. Roth, et al. A Cognitive Task Analysis for Cyber Situation Awareness. In *Proc. Human Factors and Ergonomics Society Annual Meeting September 2010 vol. 54 no. 4* pages 279-283, 2010.

Merit. Michigan Cyber Range. Merit Network, Inc, Ann Arbor, MI, 2013. http://www.merit.edu/cyberrange/.

MIT/LL and IHMC. Design Considerations for Ontology Evolution in Loosely Coupled Development Communities. MIT Lincoln Labs, Cambridge, MA, 2013.

K. Parrish. Cyber May Be Biggest Threat, Hagel Tells Troops. 2013. http://www.defense.gov/news/newsarticle.aspx?id=120178.

C. Pellerin. Cyber Tops Intel Community’s 2013 Global Threat Assessment. 2013. http://www.defense.gov/News/newsarticle.aspx?ID=119776.

S.L. Pfleeger, D.D. Caputo. Leveraging behavioral science to mitigate cyber security risk. *Computers & Security*, 31(4):597-611, 2012.

R. Pokorny. Navy awards IAI a new contract to develop Computational cyber-security Attacker/Analyst Models. Intelligent Automation, Inc., Rockville, MD, 2013. http://www.i-a-i.com/?News/2013/navy-awards-iai-a-new-contract-to-develop-computational-cyber-security-attackeranalyst-models [27]Ponemon Institute. 2013 Annual Cost of Failed Trust Report: Threats & Attacks. Ponemon Institute, 2013.

L. Pridmore. National Cyber Range: Flexible Automated Cyber Test Range (FACTR). In *Proc. Modeling and Simulation Multi-Con (MODSIM 2012)*, NDIA, 2012.

E.M. Sims, G. Taylor. Modeling Believable Virtual Humans for Interpersonal Communication. In *Proc. Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*, 2009.

R. Sweet, M. Metersky, et al. Command and Control Evaluation Workshop. In *Proc. MORS C2 MOE Workshop*, Military Operations Research Society, 1985.

M. van Lent. *Learning Task-Performance Knowledge Through Observation*. Ph.D. Thesis at University of Michigan, Department of Electrical Engineering and Computer Science, 2000.

P. Weyhrauch. MOC-WAR. 2013. http://www.navysbir.com/13_2/117.htm.