Navy Personnel Research, Studies, and Technology Division
Bureau of Naval Personnel (NPRST/BUPERS-1)

Millington, TN 38055-1000 **NPRST-TN-08-9** **September 2008**

Agent-based Approaches to Dynamic Team Simulation

Katia Sycara, Ph.D.
Carnegie Mellon University

Michael Lewis, Ph.D.
University of Pittsburgh

NPRST
research at work

NPRST-TN-08-9
September 2008

Agent-Based Approaches to Dynamic Team Simulation

Katia Sycara, Ph.D.
Carnegie Mellon University

Michael Lewis, Ph.D.
University of Pittsburgh

Reviewed by
Paul Rosenfeld, Ph.D.
Institute for Organizational Assessment

Approved and released by
David L. Alderton, Ph.D.
Director

Navy Personnel Research, Studies, and Technology (NPRST/BUPERS-1)
Navy Personnel Command
5720 Integrity Drive
Millington, TN 38055-1400
www.nprst.navy.mil

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
09-31-2008	Technical Note	1 June 2007 - 31 December 2007

4. TITLE AND SUBTITLE

Agent-based Approaches to Dynamic Team Simulation

5a. CONTRACT NUMBER

N00189-07-P-Z462

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Katia Sycara, Ph.D.
Michael Lewis, Ph.D.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Navy Personnel Research, Studies, and Technology (NPRST/BUPERS-1)
Bureau of Naval Personnel
5720 Integrity Dr.
Millington, TN 38055-1000

8. PERFORMING ORGANIZATION REPORT NUMBER

NPRST-TN-08-9

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213

University of Pittsburgh
Pittsburgh, PA 15260

10. SPONSOR/MONITOR'S ACRONYM(S)

CMU/Pitt

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

A - Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Office of Naval Research (ONR) has recently funded the Personnel Integration of Selection, Classification, Evaluations, and Surveys (PISCES) effort to develop selection, classification, assessment, assignment, and cost metrics for both individuals and teams; Total Force Assessment metrics; team configuration metrics; and tools to allow increased human resource flexibility while significantly lowering transaction costs. One of the products of the PISCES effort involves creating a virtual environment to enable team simulations, This report provides background in key areas of virtual environments and simulations, including a review of agent-based models of teamwork, discussion of the advantages and disadvantages of different levels of complexity, and using the team simulation for diagnostic purposes. Using the information provided here, decisions can be made prior to embarking on the project on the degree of sophistication of the agent reasoning, the intended functionality of the simulation (prediction of performance, selection to a team, diagnostic assessments, etc.), and the type and interface required.

15. SUBJECT TERMS

Intelligent agent, simulation, team

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Genni Arledge
UNCLASS	UNCLASS	UNCLASS	UNCLASS	61	19b. TELEPHONE NUMBER *(Include area code)* 901-874-2115 (882)

Foreword

The Department of the Navy has only recently begun to structure its human resources in an integrated, Total Force manner. To improve the Navy's ability to develop, support and maintain an integrated Total Force, the Office of Naval Research has funded the Personnel Integration of Selection, Classification, Evaluations, and Surveys (PISCES) effort. This effort encompasses a variety of goals, including the development of selection, classification, assessment, assignment, and cost metrics for both individuals and teams; Total Force assessment metrics; team configuration metrics; and tools to allow increased human resource flexibility while significantly lowering transaction costs.

As part of the PISCES effort, a virtual environment for team simulations will be created. This report provides a review of technology available to enable this effort and is meant to assist in focusing its development. The authors would like to thank Dr. Michael White and Ms Zannette Uriell for their support and guidance.

DAVID L. ALDERTON, Ph.D.
Director

Contents

List of Figures

List of Tables

Introduction

This report defines the capabilities required to develop Test Simulator (TESTOR), an experimental agent based virtual simulation for a distributed team. These capabilities are organized around the technologies of agent-based approaches, simulation, and optimization relevant to team selection and performance. The report identifies supporting capabilities necessary for specifying and capturing team performance metrics in an experimental virtual simulation environment. Four applications of the model consistent with PISCES objectives are considered for the simulation:

- Prediction of team performance
- Team selection
- Individual diagnostic assessment of teamwork
- Assessment of teamwork

The report is organized into six sections. In the first section we characterize current knowledge of teamwork and the factors that would need to be incorporated in a comprehensive simulation of team behavior. The second section reviews agent-based models of teamwork describing work involving both teamwork approaches to design of multiagent systems and agent-based representation of human behavior. The third section examines the advantages and disadvantages of agent-based modeling in the context of the complexity and richness of human teams and explores possible methods for overcoming the difficulties. The fourth section discusses issues related to predicting team performance from simulation. Section five discusses advantages and disadvantages of conventional optimization and agent-based approaches to the team selection problem. Section six explores the problems and possibilities of using virtual team simulation for diagnostic assessment of an individual Sailor's teamwork behaviors and the extension of automatic assessment to human teams.

Teams and Teamwork

Teamwork has typically (McGrath, 1964; Salas, Dickinson, Converse, & Tannenbaum, 1992) been characterized by an Input-Process-Output (I-P-O) model consisting of *inputs* such as team composition or personalities of the team members; a *process*, in which these inputs combine to determine team behavior; and *output* defined in terms of team performance or team effectiveness. Variants of this basic model such as Kozlowski and Ilgen (2006) separate the task and situation which may be expected to vary over time from more persistent characteristics such as team composition or cohesiveness that are properties of the team itself. Other authors such as Marks, Mathieu, and Zaccaro (2001) have given greater emphasis to the temporal component characterizing team processes as recurring interleaved episodes involving planning, action, and reflection and requiring explicit consideration of dynamics. A related issue involves the widely made distinction between *taskwork*, performing an individual task

within a team, and *teamwork*, skills involved in interacting with and supporting other members of a team. Crew resource management (Helmreich & Foushee, 1993) training and related approaches explicitly target the training of teamwork skills for members already proficient in taskwork. A similar progression of teamwork developing after taskwork is noted by Cooke, Salas, Kiekel, and Bell (2004) for teams learning an Unmanned Aerial Vehicle (UAV) control task. Growing evidence (Chen, Donahue, & Klimoski, 2004; Stevens & Campion, 1994; Ellis, Bell, Ployhart, Hollenbeck, & Ilgen, 2005); however, suggests that some aspects of teamwork skills can be transferred between tasks. Stevens and Campion (1999), for example found that 8 percent of the variance in supervisor's ratings of teamwork and 6 percent in ratings of overall performance was accounted for by self-reports of teamwork skills. Such reports make it reasonable to consider evaluating teamwork skills in simulation at something other than the target task.

This report will adopt the conventional I-P-O viewpoint but follow Kozlowski and Ilgen (2006) in treating task and situational demands as a special type of input. This perspective will allow us to treat team effectiveness, the objective of this effort, as a function of Sailor selection and assignment to teams, the inputs of interest.

Task Taxonomies

Taxonomies of team tasks can be divided into three general types organized by domain, task characteristics, or function. Domain based taxonomies such as Devine (2002) rely on the observation that particular domains or job categories typically involve tasks of a few predominant types. Fire fighters, for example, would be classified by Devine as belonging to a response-type workgroup and to perform proceduralized reactive tasks in uncertain environments under stressful conditions. Fast food workers, by contrast, would be classified as belonging to a service-type workgroup and would be expected to perform proceduralized reactive tasks but in a structured environment. This broad identification of task with occupational category appears well suited for selection and assignment decisions but may work less well for behavioral modeling. A doctor's duties, for example, might involve a substantial amount of paperwork in addition to the evident knowledge and skill related activities involved in surgery.

Task characteristic based taxonomies such as that of Holland (1985) account for such inconsistencies by classifying tasks into abstract categories typically derived through factor analysis. Holland classified tasks as *realistic*, *investigative*, *artistic*, *social*, *enterprising*, and *conventional*. These descriptive categories may be useful to the extent that they can be readily related to personality traits as for example in Driskell, Salas, and Hogan (1987). They also can be shared within a job category as for example a doctor who performs *investigative* (diagnosis), *realistic* (surgery), and *conventional* (record keeping) tasks in the course of his duties. These categories can again be interpreted in terms of a predominant task type for example characterizing the predominant tasks of architects as *artistic* or of clerks as *conventional.*

Steiner (1972) proposed a functional taxonomy recently adopted by Barrick, Stewart, Neubert, and Mount (1998) that characterized tasks by team process. Steiner's system classifies tasks as:

- Additive—requiring summed performance of the group (moving a table, for example)
- Compensatory—requiring individual performance to be averaged (Delphi projections, for example)
- Conjunctive—requiring adequate performance from entire team (an aircrew with pilot, navigator, and gunner, for example)
- Disjunctive—depending on the maximum performance within group (solving a puzzle, for example)

While this scheme is well adapted for assessing performance and selection it is difficult to see how complex realistic tasks can be consistently fitted within functional categories. From an agent-based modeling perspective, these functional types of effects would be expected to emerge from the execution of tasks in simulation.

Taxonomic Alternatives

Development of a new Taxonomy of Navy Teams (ATONT) is one of the precursor activities within Personnel Integration of Selection, Classification, Evaluations, and Surveys (PISCES) contributing to the development of TESTOR. At this stage it appears likely that ATONT will be domain-based and characterize dominant tasks. For agent-based modeling the crucial consideration will be the degree of constraint imposed by the chosen task(s) which will determine the capabilities required of the agent. For constrained proceduralized tasks such as interactions among an aircrew flying a supply mission, a fairly simple implementation might suffice. A loosely constrained task such as mission planning, by contrast, would require much greater sophistication and hence greater time and cost to prepare.

Team Effectiveness

Team effectiveness refers to a comprehensive assessment of success in performance. A team that accomplishes its mission within the allotted time using the allotted resources would be considered effective. Objective effectiveness of this sort might be judged in any number of ways including supervisors' ratings, or measures of productivity such as quantity or quality. Hackman (1987) maintained that team effectiveness needed to consider outcomes affecting the team itself as well as task performance and introduced *team viability* as a complementary outcome measure. Team viability referred to team members' willingness and ability to continue working together after accomplishing their task. So for example, a racing team that won a race despite antagonizing members of the pit crew, thus decreasing team viability, would be considered less effective than a team that won without such social dislocation. Similar outcomes related to the history and experience of a team are team efficacy and team potency (Gully, Incalcaterra, Joshi, & Beaubien, 2002) referring to a team's perceived capability to perform a task (efficacy) or capabilities in general (potency).

Recommendation

Team efficacy, potency, and a variety of other factors found to affect team performance form a virtuous cycle through which good performance leads to positive affect that is in turn correlated with subsequent good performance. With meta-analysis reported correlations (Gully et al., 2002) accounting for between 10 percent (potency) and 16 percent (efficacy) of observed variance in objective measures (quantity/quality) of team performance, modeling these dynamics may be a potential use for the team simulation.

Team Processes and Assessment

While team effectiveness can often be measured objectively either through standards or through reference to the performance of other teams there are situations such as a fruitless patrol for which it is difficult to define an accurate outcome measure. Software engineering researchers attempting to assess the quality of software have resolved a similar problem by assessing the quality of the process (how the software was written and checked) rather than the product (the software itself). If a relationship can be established between characteristics of the process and measurable outcomes then in situations lacking a measurable outcome, process measurements can be used as surrogates. In the study of teams there have been many attempts (Prince & Salas, 1993; Marks et al., 2001) to identify stages and behaviors in team processes, typically with reference to team effectiveness. Twenty out of 29 models reviewed by Rousseau, Aube, and Savoie (2006), for example, record *communication* as a necessary behavior, while 7 of the models reference *monitoring and back-up behaviors* as important to team effectiveness.

For construction of agent-based team simulations, models of team processes and behaviors are important for a number of purposes:

- Defining the information transformation processes (behaviors) and interactions between agents needed to simulate human teams
- Providing more sensitive measures of team performance where outcomes may be difficult to assess
- Identifying contexts and behaviors needed for an agent to interact with human team members
- Developing process measures for assessing teamwork behaviors of human interacting with team simulation
- Automated assessment of teamwork behaviors of human teams interacting through simulation

Researchers will use the teamwork process model proposed by Rousseau et al. (2006) as an integration of 27 earlier models (8 of them listing Eduardo Salas as an author) to illustrate processes needing inclusion in a comprehensive team simulation.

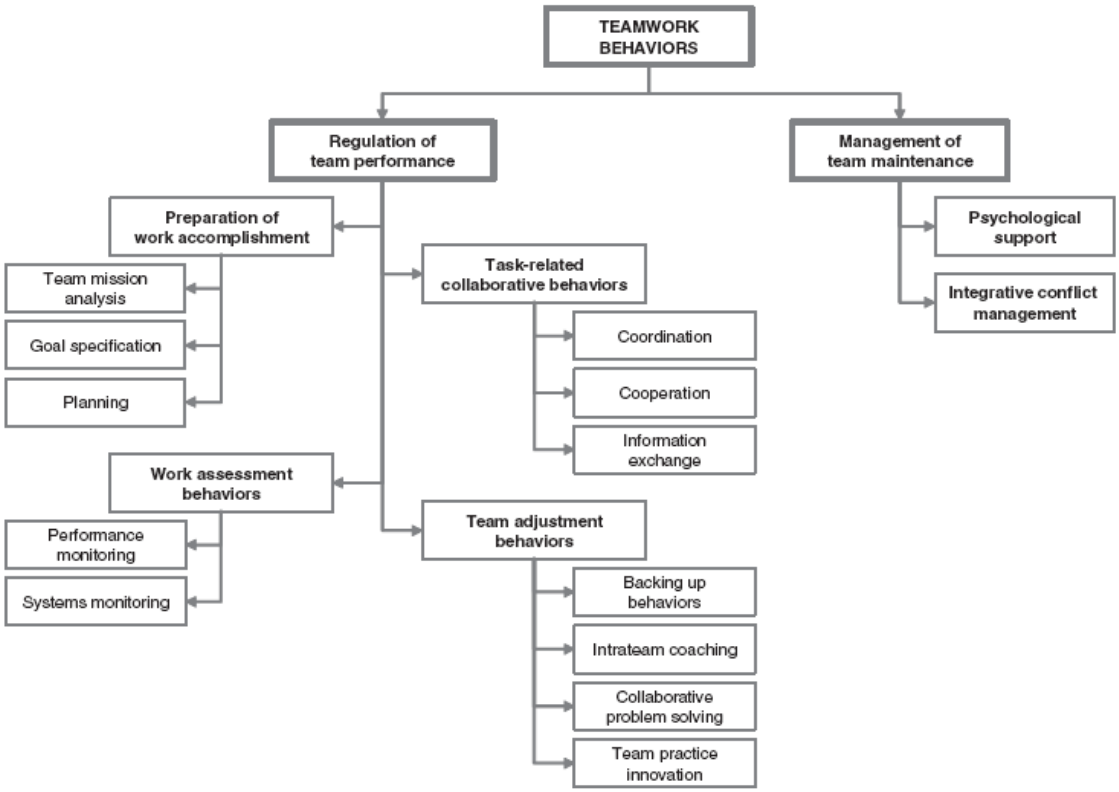

Figure 1. Schematic representation of the hierarchical conceptual structure of teamwork behaviors from Rousseau, V., Aubé, C. and Savoie, A. (2006).

The approach will be to identify subsets of these behaviors that tend to co-occur in real tasks. If a group of operationally significant tasks can be performed using a limited subset of teamwork behaviors then an agent-based model incorporating only this subset of behaviors would be sufficient for modeling this group of tasks. Figure 1 presents an ontology of teamwork behaviors. The first branching distinguishes between behaviors whose purpose is preserving the integrity and effectiveness of the team, labeled *Management of team maintenance*. The other branch labeled *Regulation of team performance* contains behaviors needed for task performance. These are further divided among behaviors involved in planning, *Preparation of team performance;* performing the task, *Task-related collaborative behaviors;* monitoring, *Work assessment behaviors;* and adaptation, *Team adjustment behaviors*. As Figure 1 illustrates, there are distinct sets of processes that may be called upon for different types of tasks. *Task-related collaborative behaviors, Work assessment behaviors*, and *Team adjustment behaviors* would be needed to simulate command and control many execution-oriented military tasks. The behaviors involved such as *information exchange*, *performance monitoring*, and *backing-up behaviors* could be specified fairly concretely and implemented as agent rules. Incorporating planning, collaborative problem solving, and other more abstract processes into an agent model would require a more complex architecture and execution process such as the hierarchical task network (HTN) planner used in RETSINA (Sycara, Paolucci, Giampapa & van Velsen, 2001). Adding *Management of team maintenance* functions would require an additional level of complexity to accommodate conflicting goals among agents and the need for explicit coordination and negotiation mechanisms. What is significant about models of team

process is that they can be assembled to accomplish tasks in a way that varies levels of complexity and that many of the tasks likely to be of most interest to the Navy (e.g.; structured well practiced tasks), can be accommodated by the simpler models.

For team members to be modeled by agents will require concrete individual I-P-O specifications as well as description of the processes through which they interact. While sensory inputs can be derived from descriptions of the task and environment, defining agent processes will require considering what team members know and think. Within the teamwork literature these skills can be described by knowledge and skills (Hackman 1992) characterizing taskwork (what to do) and teamwork (how to interact with other agents) and what is referred to as transactive memory, knowledge of how information is distributed within the team.

Criteria for Assessing Quality of Process Performance

Sensing

- Accurate detection of all available information
- Correct interpretation (attachment of correct meaning) of all detected information, to include appropriate weighing of its importance
- Accurate discrimination between relevant and irrelevant information
- Attempts to obtain information are relevant to mission, task, or problem
- Sensing activities are timely in relation to information requirements and the tactical situation of the moment
- Internal processing and recording of information provides ready availability to users

Communicating Information

- Accuracy of transmission of available information
- Sufficiently complete to transmit full and accurate understanding to receivers of communications
- Timeliness appropriate to unit requirements
- Correct choice of recipients: everyone who needs information receives it
- Whether message should have been communicated

Decision Making

- Adequacy: Was the decision adequately correct in view of circumstances and information available to the decision maker?
- Appropriateness: Was the decision timely in view of the information available to the decision maker?
- Completeness: Did the decision take into account all or most contingencies, alternatives, and possibilities?

Stabilizing

- Adequacy: Action is correct in view of the operational situation and conditions that the action is intended to change or overcome
- Appropriateness: Timing is appropriate in view of the situation, conditions, and intended effects. Choice of target of the action is appropriate
- Completeness: Action fully meets the requirements of the situation

Communicating implementation

- Accuracy of transmission of instructions
- Sufficient completeness to transmit adequate and full understanding of actions required
- Timely transmission in view of both available information and the action requirements of the participants
- Transmission to appropriate recipients
- "Discussion or interpretation" is efficient, relevant, and achieves its purpose
- Whether message should have been communicated

Coping actions

- Correctness of actions in view of both the current operational circumstances and the decision or order from which the action derives
- Timeliness of the action in view of both operational circumstances and the decision or order from which the action derives
- Correctness of choice of target of the action

Feedback

- Correctness of the decision and action to obtain feedback in view of operational circumstances, the preceding actions whose results are being evaluated, and current information requirements
- Timeliness of the feedback decision and action
- Correctness of choice of target(s) of the action
- Appropriate use of feedback information in new actions, decisions, and plans

Note. From *Battle Staff Integration*, by J. A. Olmstead, 1992 (IDA Paper P-2560), Gov. Rep., Alexandria, VA: Institute for Defense Analysis.

Figure 2. Reprinted from Millitello, Kyne, Klein, Getchell, & Thordsen (1999).

Although Rousseau et al.'s models categorize behaviors in a generally prescriptive way implying that there should be behaviors for coordinating, communicating, backing up, etc. they do not provide an instrument for classifying an observed team process as effective or ineffective. Figure 2 shows one such attempt consistent with the studies contributing to Rousseau's model to provide criteria for assessing teamwork process (Olmstead, 1992 reprinted from Milltello, Kyne, Klein, Getchell, & Thordsen, 1999). As with Crew Resource Management (CRM) training and much of the research directed by Salas under the TADMUS (Team Decision Making Under Stress) program, some of the most easily observable process characteristics are found to characterize high performance teams, typically teams working in high stress/high consequence settings such air crews, operating rooms, or the battlefield. Examination of this list suggests that with appropriate choice of task and operationalization of a subset of criteria it should be possible to automate the assessment of human teamwork process performance. In particular, because these measures of process can be associated with individual team members rather than the performance of the team as a whole it could provide an instrument to look inside a team allowing for individual assessments independent of overall team performance.

Attributes Influencing Performance

While the I-P-O relation between team members and the environment produce the behaviors researchers observed in real teams and hope to model in simulation there are a variety of characteristics of individuals and teams that moderate this relation. Because moderating attributes are things that could be measured in individuals or teams and used in making selection decisions, they are a desirable part of the team simulation.

Individual Differences

GMA. General mental ability of team members has been found to correlate with team performance in as varied areas as crews of soldiers (Tziner & Eden, 1985), systems analysts (Hill 1982), and supervisor's ratings on technical skills, teamwork, and team performance in production lines (Stevens & Campion, 1994). Stevens and Campion additionally found correlations of .36, .23, and .29 between mean scores on an aptitude test and supervisors' ratings for a team's technical skills, teamwork, and performance suggesting that average GMA may account for approximately 10 percent of variance in team performance. Williams and Sternberg (1988) additionally found correlations with a team's highest individual intelligence score suggesting the potential usefulness of notions from Steiner's (1972) functional task taxonomy in modeling disjunctive tasks.

Personality. Although there is substantial evidence (Barrick & Mount, 1991) of association between the 5-factor personality model and individual performance there is less direct evidence for teams. Hough (1992), for example, found that ratings on conscientiousness, emotional stability, and agreeableness were correlated with ratings of cooperativeness with coworkers and team members, but did not include measures of team performance in his analysis. Peeters, Rutte, Tuijl, and Reymen (2006) who found agreeableness and emotional stability positively related to satisfaction with the team make similar conjectures about the relation between agreeableness and teamwork. In studies linking personality to team characteristics Schneider, White, and Paul (1998) again found agreeableness to account for 8 percent of the variance in measures of fit to an organization. There appears to be better evidence for balance among personality types as a determinant of team effectiveness. Barry and Stewart (1997), for example, found a curvilinear relation between the number of extraverted team members and team effectiveness, with teams with too few or too many extraverts performing less well. Stewart and Barrick (2004) found another compositional effect in which a single member low on agreeableness or emotional stability was sufficient to degrade team effectiveness. Peeters et al. (2006) found a positive correlation between satisfaction and dissimilarity in conscientiousness as well as a negative relation for dissimilarity in extraversion for members low on the trait. As Kozlowski and Ilgen (2006) point out "well-developed theoretical models are needed to help specify complex patterns of composition." Such development would be needed before multi-agent compositional effects such as those related to dissimilarity or emotional stability could be modeled within an agent-based simulation.

Task Knowledge and Skills. While the teamwork literature focuses on teamwork and team skills real tasks depend largely on team members' abilities to perform their assigned taskwork. Composing teams in terms of requisite skills is a well-known problem readily handled by operations research techniques that will be reviewed later. Task skills must be incorporated into any agent-based model because of their dominant effects on team performance. A plane could be flown by a dull and neurotic pilot, for example, but not by an intelligent and agreeable flight crew that did not include a pilot.

Team Attributes

In addition to effects associated with individual characteristics and the composition of teams there are several widely reported team level characteristics that have been shown to be related to team effectiveness.

Cohesiveness. Team cohesion refers to the degree to which team members report identifying with the team and team goals and has been widely studied. Whether considered at the individual or team level, cohesion has been consistently shown to improve both team processes and team performance. In meta-analyses by Gully, Divine, and Whitney, (1995) and Beal, Cohen, Burke, and McLendon (2003), cohesion was found to have an effect size of approximately r = .3-.4 with greater effects noted at the team level and greater effects for teamwork behaviors than outcomes. Cohesion was also found to be a greater factor accounting for almost 22 percent of the variance (Gully et al. 1995) for highly interdependent tasks. Reports based on field interviews such as Shils and Janowitz (1948) classic study of the German Wehrmacht frequently cast cohesion in the even stronger role of serving as a buffer against otherwise intolerable stresses in combat. Griffith (1997) and Griffith and Vaitkus (2000) claim this to be its primary role and propose models in which cohesion serves as a moderator or mediator rather than a main effect on performance. As the data from studies included in the earlier meta-analyses measure performance primarily through self-reports, ratings on exercises, and other noncombat settings the importance of cohesion to performance in combat is likely underestimated. The meta-analyses, however, substantiate a robust measured relation between cohesion and team and individual performance that make a meaningful contribution whether directly or indirectly to prediction. Another feature that may bear incorporation into later models is predictable dynamic behavior. For manpower intensive units in the military, Siebold (2007), for example, reports that cohesion follows a predictable U-shaped curve, starting out at a high level, beginning to decline at approximately three months, bottoming out at approximately a year, then increasing from there to regain approximately half its initial level.

Climate. Organizational climate has been studied widely for almost 70 years and consistently shown to relate to team behavior and outcomes. Schneider and Bowen (1985) showed that a shared climate involving service predicted customers' satisfaction with their bank branch while Hofmann and Stetzer (1996) found a team climate for safety predicted safety-related behaviors and actual accident rates in a chemical plant. In a recent meta-analysis Carr, Schmidt, Ford, and DeShon (2003) estimated correlations of r = .09 and r = .05 between affective and instrumental aspects of climate and individual performance.

Efficacy and Potency. Mentioned earlier in the section on team effectiveness, team efficacy, the team's belief in its abilities to perform a task and team potency, the team's confidence in its general abilities, are additional team-level constructs demonstrated to enhance team performance.

A Model of Teamwork Incorporating Attributes

This brief survey can be summarized in the schematic model shown in Figure 3. The factors identified as influencing teamwork do not actually determine what team members do but rather how well they work together. In this model a normative team (without individual or group characteristics) interacts through a normative process (a work flow specifying conditions and actions) with its task and environment. An agent-based model of a work flow of this sort can be readily programmed. This interaction is moderated by individual differences, team composition, and team attributes. If the magnitudes of effect estimated in this section were additive such a model might account for up to three-quarters of the variance in performance among teams. Such a result, however, is extremely unlikely because constructs such as team efficacy, cohesiveness, individual agreeability, and general mental ability are almost surely highly correlated and likely to interact over time in complex ways. Constructing an accurate model predicting differential team behavior from such theory and data would require describing these relations precisely.

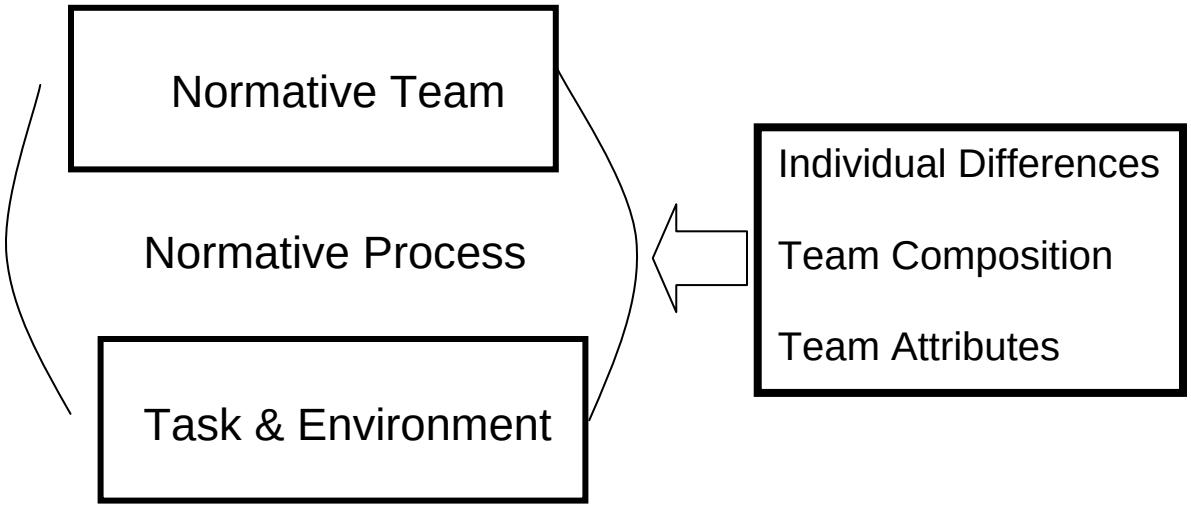

Figure 3. Teamwork model: Behavior of normative model is moderated.

Agent Models for Team Simulation

The study of autonomous agents and multi-agent systems centers around the concept of an agent. An agent is an information processing system that can receive inputs from its environment and act in turn upon that environment. A rational agent is one that acts so as to optimize some performance measure. Because the capacity of a

rational agent is limited by its knowledge, its computing resources, and its perspective an agent can exhibit only bounded rationality (Simon, 1957). Numerous works in artificial intelligence (AI) research try to formalize a logical axiomatization for rational agents (see Wooldridge & Jennings [1995] for a review). This axiomatization is accomplished by formalizing a model for agent behavior in terms of beliefs, desires, goals, and so on. These works are known as belief-desire- intention (BDI) systems (Rao & Georgeff, 1991; Shoham, 1993). An agent that has a BDI-type architecture has also been called *deliberative*. This means that its actions are determined by matching beliefs to desires to determine intentions rather than simply matching inputs to predetermined actions as done with "if-then" production rules. While early AI research attempted to develop systems realized as single precocious agents, subsequent research has led to the development of multiagent systems (MAS) in which intelligence is modularized. Making such systems work required developing theories about the basic requirements for coordinated and cooperative behavior. Two dominant perspectives are *joint intention* (Cohen & Levesque, 1990) and *SharedPlans* (Grosz & Kraus 1996). *Joint intention* holds that teamwork requires maintaining commitment to common goals and requires communication for grounding shared beliefs about the state of the task and changing circumstances. According to *shared-plans,* agents must have a common goal, agree on the recipe for accomplishing that goal, and accept assigned roles for working toward that goal.

While theory and research involving agents originated in the distributed AI community the current field has been greatly expanded to include the study of markets and auctions by economists, the behavior of schools of fish or swarms of robots by biologists and control theorists, the interactions of self-interested agents by game theorists, and many other application areas. This review will focus on forms of MAS that include mechanisms most likely to characterize the behavior of Navy teams. These mechanisms include: sharing of goals, sharing of plans, and assignment of roles.

This section introduces the RETSINA multiagent architecture as an example of a MAS with facilities for modeling all of the needed mechanisms. A less general approach to teamwork pointing out advantages and disadvantages is described. Distinctions and difficulties in modeling naturally occurring teamwork phenomena using variable-based or agent-based models are also discussed. Applications of agent-based models to modeling human behavior and discussion of the issues likely to arise in modeling Sailor teams are then reviewed.

RETSINA: An Example of a Full Featured MAS

Extending *joint intentions* and *shared-plans* that assume a closed world and small homogeneous teams, RETSINA provides a multiagent infrastructure for finding, assembling, and coordinating teams of agents to accomplish specified goals. RETSINA has been developed under the following assumptions: (a) the agent environment is open and unpredictable (i.e., agents may appear and disappear dynamically), (b) agents are developed for a variety of tasks by different developers that do not collaborate with one another, (c) agents are heterogeneous and could reside in different machines distributed across networks, and (d) agents can have partially replicated functionality and can incorporate models of tasks at different levels of decomposition and abstraction. For

example, there can be a single agent that provides all kinds of weather information (including barometric pressure, wind direction etc.) for all cities in the world. On the other hand, there could also be weather agents that provide only temperature. Alternatively, there can be an agent that provides radar operator functionality, and agents that provide only target tracking functionality (a subtask of the radar operator task) for a particular environment. These agents could vary in fidelity to the task constraints (e.g., the target tracking agent could operate at a more refined resolution level for tracking).

To be an effective team member, besides doing its own task well, an agent must be able to receive tasks and goals from other (appropriate) team members, be able to communicate the results of its own problem solving activities to appropriate participants, monitor team activity and delegate tasks to other team members. A pre-requisite for an agent to perform effective task delegation is to know (a) which tasks and actions it can perform itself, (b) which of its own goals entail actions that can be performed by others, and (c) who can perform a given task. The individual agent architecture (shown in Figure 4) that was developed (Sycara et al., 2001) includes abilities of agents to send messages to one another (RETSINA agents communicate using Knowledge Query and Manipulation Language [KQML]), declarative representation of agent goals and planning mechanisms for fulfilling these goals. Therefore, an agent is aware of the objectives it can plan for and the tasks it can perform. In addition, the planning mechanism allows an agent to reason about actions that it cannot perform itself but which should be delegated to other agents. To do so, an agent needs ways to find out the capabilities of other team members (i.e., what tasks other agents can perform). As shown in Figure 4, each agent has a communications module, which is responsible for interactions and the exchange of messages with other agents. These messages could contain new objectives from other agents or from the environment. The communicator uses the input/output message queue to modify the agent's set of high-level objectives in its knowledge store. The planner module uses the objectives and a plan library of pre-specified plan fragments. The planner composes these plan fragments to construct alternative possible plans for the agent, stored as task structures. The scheduler module uses the task structures determined by the planner module to create a schedule of primitive actions for execution that the agent can then execute. The execution monitor module monitors action execution in the operating environment and suggests repairs if actions fail.

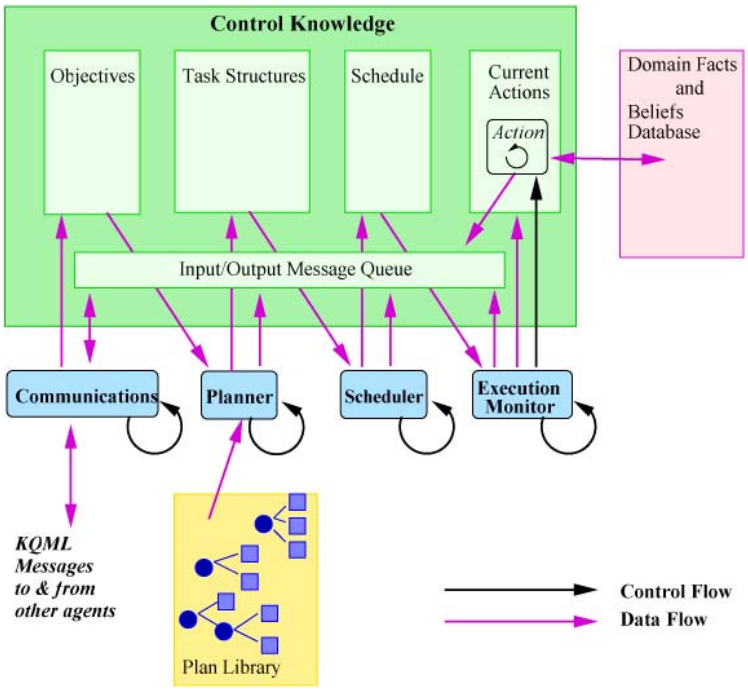

Figure 4: Individual RETSINA agent.

The four modules operate in parallel as multi-threaded code. Thus, the agent can receive messages from other agents through the communicator module while the planning module simultaneously constructs plans. In this way, an agent can interleave deliberative planning with information gathering and execution monitoring, an important capability in dynamically changing environments.

In addition an agent has a knowledge store which consists of a goal stack, where incoming or internally generated goals are stored; and a task database, where task fragments relevant to the agent's functionality are stored and reused to construct plans. The agent model also contains an additional belief database that stores the current beliefs of the agent that can change due to evolving situation changes or due to agent-internal processing.

Since every agent can both plan and execute action sequences, the above architecture enables deliberation and reaction to the environment to be performed as needed at every stage of the task decomposition. In addition, the system does not impose on agents a particular granularity of task decomposition. Based on the environment and on constraints passed down from other agents, a agent can plan and choose the best course of action. This enables it to flexibly and dynamically adapt to changes in the environment including changes in the goals and intentions of its teammates.

Specialized Models of Teamwork

To implement a software system, we must select coordination and communication mechanisms that the agents can use. For some domains, simple pre-arranged coordination schemes like the locker-room agreement (Stone & Veloso, 1999) in which the teams execute pre-selected plans after observing an environmental trigger are adequate. Although this coordination model has been successful in the Robocup domain, the locker-room agreement breaks down when there is ambiguity about what has been observed; what happens when one agent believes that an event trigger has occurred but another agent missed seeing it? The TEAMCORE framework (Tambe 1997) recently reimplemented in the Machinetta system (Scerri, Pynadath, Schurr, Farinelli, Gandhe & Tambe, 2004) was designed to address this problem by executing "canned plans" more flexibly. TEAMCORE agents reason explicitly about goal commitment, information sharing, and selective communication to coordinate their actions. The behavior of these agents is based on team oriented plans (TOPs), which describe joint activities to be performed in terms of the individual roles to be performed and any constraints between those roles. TOPs are instantiated dynamically from TOP templates at runtime when preconditions associated with the templates are filled. A team of Unmanned Combat Air Vehicles (UCAVs), for example, might execute a variety of attack TOPs.

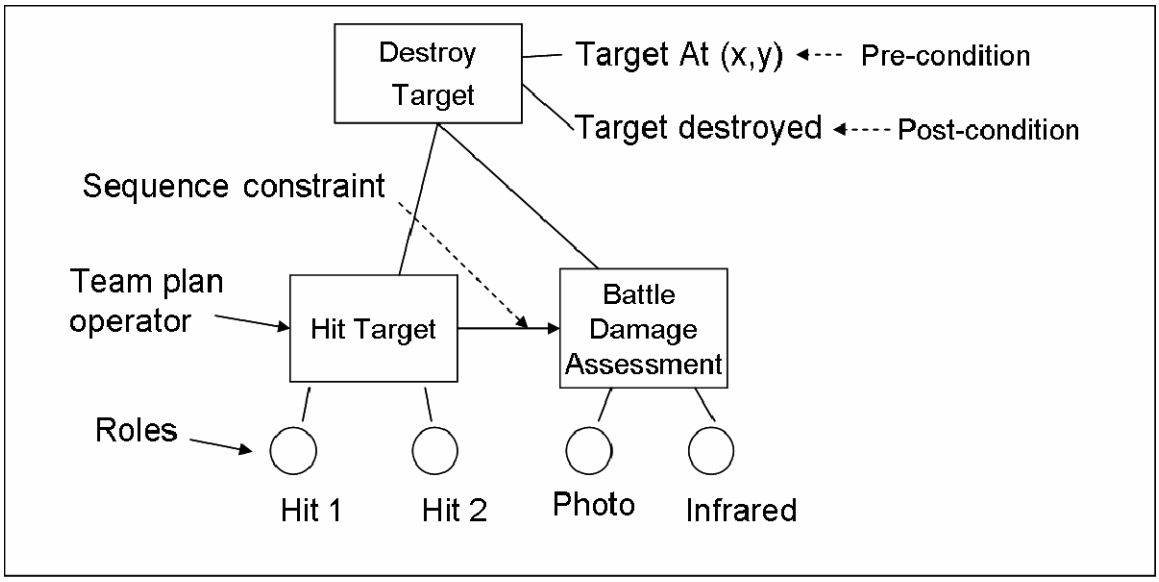

Figure 5. A TOP for attack and BDA.

When a UCAV identifies a target in an open area it might instantiate a simple attack TOP and send out a request to fill second attacker and Battle Damage Assessment (BDA) roles. After the roles are filled two UCAVs attack the target and the third follows to record the damage (Figure 5). Another UCAV spotting a convoy of trucks near cover might instantiate a more complex simultaneous attack plan requiring filling multiple attacker roles in order that they might attack together to catch the convoy in the open. Constraints between these roles will specify interactions such as required execution

ordering and whether one role can be performed if another is not currently being performed. Because behavior of agents in this scheme is much more constrained than in the more general RETSINA architecture, programming and simulating team behavior would be easier. Coordinated attack scenarios for a human team could be constructed in the same way as the UCAV example above. In terms of Rousseau et al.'s (2006) teamwork process model this approach could characterize team behavior for codified procedural tasks involving only work assessment or task-related collaborative behaviors. To extend modeling to less constrained tasks would require choosing a less constrained architecture.

Agent-based Modeling (ABM) vs. Variable-based Modeling (VBM)

MAS as discussed to this point have been systems designed by computer scientists to solve problems and perform tasks. The insights they have revealed involve things such as the necessity of communication, modeling of beliefs, etc. for coordination among agents to occur. While many of the constructs in MAS were clearly inspired by human behavior (e.g.; the BDI formulation is often referred to as *folk psychology)*, there is no guarantee that the resulting MAS will model human behavior. Social scientists and economists have approached the problem from the other direction constructing MASs with the particular goal of simulating key theoretical elements of some social or psychological process (Smith & Conrey, 2007; Parunak, Savit, & Riolo, 1998). Exemplars of this approach include Schelling (1971) who demonstrated that agents following a simple decision rule of moving to avoid being in a minority of < 30 percent resulted in nearly complete segregation of neighborhoods in a 2-dimensional grid. Kalick and Hamilton (1986) conducted a similarly counterintuitive demonstration showing that the finding that people tend to pair with partners of approximately the same attractiveness (r = .6) was more consistent with a population in which each agent seeks to maximize its partner's attractiveness than one in which agents actually preferred partners of comparable attractiveness.

The Kalick and Hamilton study illustrates the basic paradigm of agent-based modeling in that data are modeled at both the micro and macro level. The micro level of the model is captured by the mate-choice rules of the agents. This rule was hypothesized on the basis of studies such as Walster, Aronson, Abrahams, and Rottmann, (1966) which found that students preferred more attractive dates rather than those of more nearly the same attractiveness. At the macro level the model produces a correlation between attractiveness of mates that is closer to that actually observed than the alternative, the correlation produced in a population seeking mates of their own level of attractiveness.

A primary distinction between ABM and conventional VBM is the way in which macro level behavior is predicted. For VBM, especially parametric models, there are principled ways of attributing performance to particular constituents of the model and assigning significance levels to them. In a regression model, for example, the variables with the greatest contribution to prediction are typically entered first with additional variables judged and entered based on their contributions to explained variance. This transparency allows the modeler to choose a model that fits but does not over fit the data. ABM offers no such protections. The Schelling (1971) and Kalick and Hamilton

(1986) models both pass a face validity test for parsimony and hence are compelling. Had the dating example included measures of personality, socio-economic status, and education level it would almost certainly more closely approximate the gamut of factors that enter into real dating decisions but the strength of evidence for the role of attractiveness would likely be lost. Because of this need for parsimony and difficulty in validation, ABM has been used primarily as a confirmatory method to demonstrate the feasibility of producing an observed result from a hypothesized mechanism.

Data Derived Cognitive Models of Human Behavior

While both earlier examples involve agent-based models of humans, the agents and their behaviors themselves are quite abstract and make no attempt to characterize humans or their environment in any detailed way. The agents in the segregation study for example are one of two colors (red/green) and allowed to move about a grid from one node to another. The dating agents were assigned numbers 1–10 and proposed/accepted offers with the associated probability (.10–1.0). While this degree of abstraction was useful for demonstrating the *feasibility* of an observed outcome resulting from a behavioral mechanism it lacks the precise specification of behaviors that would be desirable for models that are intended to be predictive, perhaps even in the absence of outcome data for validation. The data-derived approach bases its claim on the construct validity of its data based micro model. If outcomes can be shown to match (macro validity), the agreement is interpreted as supporting the micro model itself rather than just its feasibility. This approach is basically deductive rather than inductive. The micro model is presumed to simulate behavioral processes in the same way that a Newtonian model of a pulley system might predict the movements and locations of the weights. Because such models of human behavior are inherently complex, parsimony cannot be claimed to justify validity and matching outputs typically involves substantial tuning. The following subsection presents well-known data-driven models that aim to match human cognitive processes. All but one of these models; however, are for individual tasks and performance and say nothing about teamwork.

ACT-R

Data-derived models have most often been used to characterize behavior at simple tasks devoting elaborate detail to cognitive processes involving perception and memory. John Anderson's ACT-R cognitive model (Anderson & Lebiere, 1998) is the most thoroughly developed model within this group. ACT-R has two types of modules: perceptual-motor modules that provide an interface between ACT-R and its simulated environment and memory modules that contain beliefs (declarative memory) or production rules (procedural memory). Data resides in buffers simulating brain areas that are searched for matches with production rules to fire. There are typically repeated modifications of buffer contents with occasional firings leading to actions or collections of input from the perceptual-motor modules. ACT-R was developed for and excels at predicting performance at controlled tasks of the sort found in psychological

laboratories. It accurately simulates performance at memory and learning tasks and more recently predicts areas of cortical activation. ACT-R is clearly the best cognitive simulation for behavior that occurs within short (.05–1 sec) time spans and for modeling effects that depend on details of memorial or perceptual processing.

Soar

While ACT-R attempts to model cognition from a structural perspective, Soar (Rosenbloom, Laird, & Newell, 1993), takes a functional view. Based on Allen Newell's (1990) unified theory of cognition, Soar incorporates learning through chunking in such a way that every decision is based on the current interpretation of sensory data, the contents of working memory created by prior problem solving, and any relevant knowledge retrieved from long-term memory. Soar is less faithful to psychological peculiarities of human cognition and more focused on abstract learning mechanisms based on problem spaces that allow "intelligent" behavior to emerge from experience. This detachment from the "hardware" allows Soar to model human behavior at a greater level of granularity. So Soar could be expected to do things such as model learning through analogy or generalizing a solution to a new problem. In some applications Soar seems to serve more as an expert systems shell than a cognitive model.

COGNET iGEN

COGNET/iGEN (Cognet, 2008), the primary product of Wayne Zachary's CHI Systems, is an expert system shell designed to incorporate some aspects of cognitive models. As such, it is much easier to insert into relatively complex scenarios than either ACT-R or Soar. COGNET basically does what it is told, so it is possible to program complex and sophisticated behaviors without having to learn them (Soar) or decompose them into "bit-level" processes (ACT-R). By limiting itself to modeling *expert performance*, considered to be "rich and highly compiled knowledge structures that have chunked many lower level productions.." (Zachery, Santarelli, Ryder, Stokes, and Scolaro, 2001), it can be programmed and run as a production system using a blackboard as a stand-in for working memory. Human frailty is added through incorporating factors limiting performance such as visual acuity or sensory noise to produce what Zachery refers to as a "performance model" representing both expertise and limitations in human expert performance.

Micro Saint / IPME

Micro Saint/IPME (Microsaint, 2008) is a product of Micro Analysis and Design, now a division of Alion Science and Technology. Micro Saint harks back to the early crew modeling simulation SAINT (Siegel & Wolf, 1967). Their approach to operator modeling was essentially a queuing simulation. By simulating the arrival and disposal of tasks by members of an aircrew the modelers hoped to identify aspects of task design or physical layout that might lead to the build up of more queued tasks than a crewmember could perform within an allotted time. In its modern form Micro Saint provides a general discrete event simulator with a task network model for human actions. The task

networks are similar to COGNET's "expertise model" but more rigid, because they lack a blackboard and follow programmed workflows. Individual differences such as level of training can be programmed directly into the crewmember models. Finally, Performance Shaping Functions (PSFs, the functional expression of performance shaping factors such as stress or fatigue) can be defined at the task level to alter the probability of success/failure in response to changes in the environment.

Figure 6 shows an example of an unusual performance shaping function from Swain and Guttman (1983) used in probabilistic risk assessment where the approach was first developed. This PSF raises the probability of human error to 1.0 immediately following a nuclear accident declining to 0.1 only after a half hour. Two hours after the accident has occurred PSF probabilities decline to the point that the probability of error is once more being determined by the task being performed rather than the PSF. A variety of mathematical approaches (Hollnagel, 2000) have been used to moderate predicted behavior using PSFs, but all share the logic of perturbing a normative response to reflect changes in context.

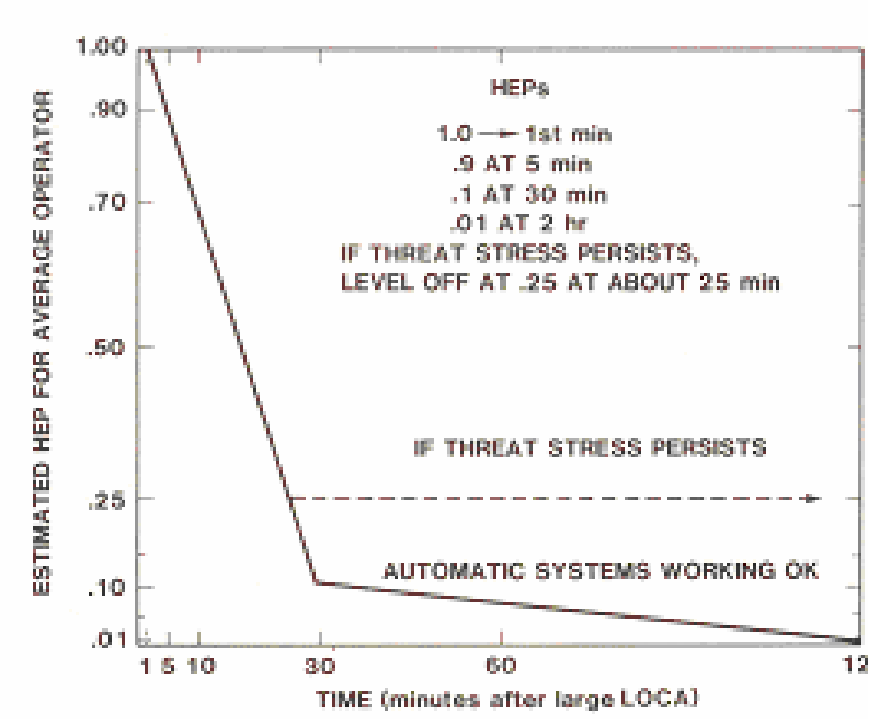

Figure 6. PSF for a large scale Loss of Cooling Accident (LOCA) from NUREG/1278 Handbook of Human Reliability Analysis.

Performance Comparison of Cognitive Models

Table 1 shows a variety of other cognitive modeling systems of varying degrees of fidelity and granularity. From 1999–2004, the Air Force Research Laboratory Human Effectiveness Directorate (AFRL/HE) sponsored an Agent-Based Modeling and Behavior Representation AMBR program (Deutsch et al., 2004) to compare and evaluate available models. The major contenders discussed earlier (ACT-R,

COGNET/iGen, EPIC-Soar along with DCOG [an AFRL-developed model]) were evaluated. The tasks that were compared with the performance of human participants were:

- Much simplified Air Traffic Control (ATC) task using either a textual or a GUI interface
- Concept learning task involving spatial relations and embedded in the ATC display
- Transfer of training test for learned concept

Across the tests COGNET/iGEN more closely approximated human performance than other models. Figure 7 reprinted from Tenney and Spector (2001) shows comparisons for penalties and times as a function of workload.

Table 1
Human behavior representation architectures available for use

ARCHITECTURE	Reference URL
ACT-R	http://act-r.psy.cmu.edu/
ART	http://web.mst.edu/~tauritzd/art/
Brahms	http://www.agentisolutions.com/home.htm
CHREST	http://www.psyc.nott.ac.uk/research/credit/projects/CHREST
Clarion	http://www.cogsci.rpi.edu/~rsun/clarion.html
Cogent	http://cogent.psyc.bbk.ac.uk
COGNET/iGEN	http://www.chisystems.com
D-OMAR	http://omar.bbn.com/
Emergent	http://www.cnbc.cmu.edu/Resources/PDP++//PDP++.html
EPAM	http://www.pahomeschoolers.com/epam/
EPIC	http://www.umich.edu/-bcalab/epic.html (no download)
MicroPsi	http://www.micropsi.org/project.php
Micro Saint, IPME	http://www.maad.com/MaadWeb/products/prodma.htm
MIDAS	http://human-factors.arc.nasa.gov/dev/www-midas/index.html (no download)
SimAgent	http://www.cs.bham.ac.uk/research/projects/poplog/packages/simagent.html
Soar	http://www.soartechnology.com

Adapted from Deutsch, Pew, Tenney, Diller, Godfrey, Spector, Benyo, and Date (2004), Table 1 organizes many of the currently available Human Behavior Representation Architectures. URLs valid as of 12/21/2007.

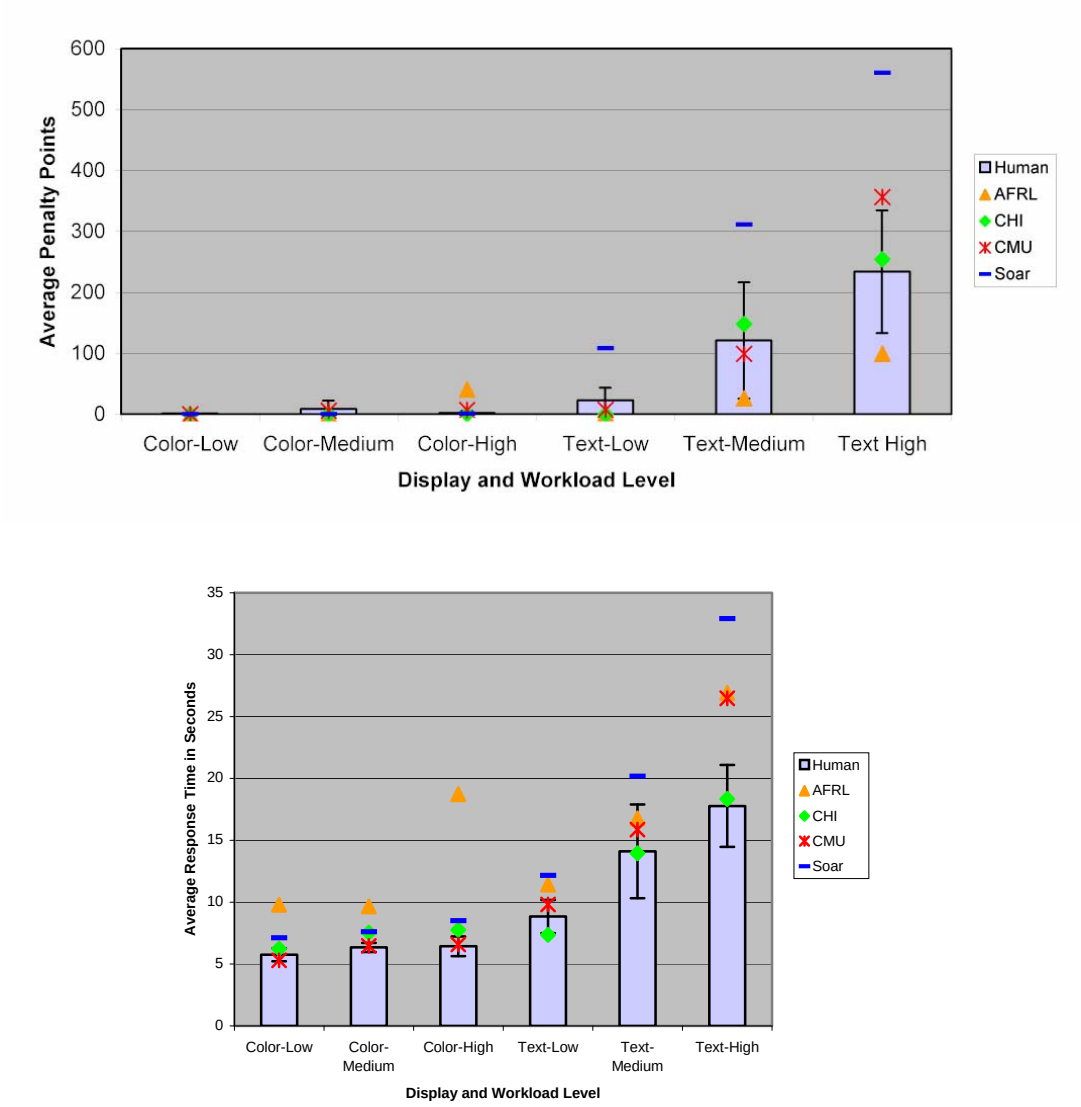

Figure 7. Display and workload level for penalties and average response times reprinted from Tenney & Spector (2001) (AFRL is DCOG, CHI is COGNET/iGEN, CMU is ACT-R, and Soar is EPIC-Soar).

This result should not be surprising given that COGNET/iGEN was developed expressly to model expert performance at procedural reactive tasks at this time scale. ACT-R which devotes greatest effort to atomic cognitive processes, faces difficulties in modeling something as complex as the ATC task at such great detail. By explicitly programming limitations for the test task to produce a *performance* model from its *expertise* model COGNET maximizes its opportunity to match human performance at any particular task but this process would need to be repeated for each new task. Although it was not tested in this program, Micro Saint/IPME, which also models at the task level might be expected to produce similar performance but be even less generalizable.

Simulated Humans

While data-derived cognitive models attempt to model the mechanisms generating human behavior, *simulated humans* are models designed to convey the *appearance* of human behavior. Since the advent of sophisticated computer games and military simulations, especially those using semi-autonomous forces (ModSAF, JSAF, OTB, OneSAF, etc.), in the 1990s there has been a need to supply believable opponents and other actors. This is often very difficult because of the complexity of the environments. In computer games, for example, simulated entities are often limited to moving along arcs between nodes of a graph with their movements restricted to a set of preprogrammed animations.

Efforts to make behaviors more believable may consist of things such as adding randomness to paths or varying an actor's speed. At the other end of the spectrum some games have become quite sophisticated with bots (agents within the game) that cooperate in attacks. Much of the research in this area is presented at a yearly conference originally called Computer Generated Forces and Behavior Representation (CGF-BM) and renamed Behavior Representation in Modeling and Simulation (BRIMS) in 2003.

Work in this area is varied but its flavor is probably best characterized by looking at several studies. Again, except for TEAMCORE, these are models of individuals and say nothing about social teamwork. As might be expected, several of the cognitive models introduced earlier have been used in this area as well. Best, Lebiere, and Scarpinatto (2002), for example, use ACT-R to model synthetic MOUT (military operations on urban terrain) opponents. A major difficulty and a substantial portion of their paper is devoted to the problem of extracting information from the game environment in a form usable by their model. Because game programmers rely on artifices such as labeling a node as an "ambush point" to avoid having to perceive the environment, data from function calls available to the applications programming interface (API) had to be used for algorithms, in this case based on Hough transforms and binary space partitioning (BSP) trees, to extract information in usable form for ACT-R. In the end agents were supplied with productions such as "If there is an enemy in sight and there is no escape route then shoot at the enemy" to produce MOUT opponent behavior.

Tambe's (1997) TEAMCORE teamwork approach introduced earlier was originally presented by Hill, Chen, Gratch, Rosenbloom, and Tambe (1997) as an application in Soar to provide CGF's (helicopters) for ModSAF. A more typical cgf team application for ModSAF is described by Reece (2003) who modeled team behavior as a hierarchy of tasks distributed over unit leaders and unit members. An A* search[1] algorithm over a 2-dimensional regular grid and a topological map was then used to produce a plan in the form of a series of waypoints annotated with posture and speed changes for the individual vehicles to follow. As these examples suggest, as complexity increases both in interacting with the simulation and in finding solutions demands of the task, heuristics, and plausibility tend to replace cognitive fidelity as the objective in modeling.

[1] A* is a best-first, graph search algorithm that finds the least-cost path from an initial node to a goal node.

More recently there has been a shift in emphasis toward social and cultural plausibility of simulated humans. A body of work from the University of Southern California is typified by the ELECT BiLat simulation (Hill, Belanich, et al., 2006). This training simulation generates culturally appropriate synthetic characters that interact with trainees using both verbal and non-verbal behaviors to train students in culturally appropriate/effective modes of interaction. While there is no pretense that the synthetic character models in an accurate way the human it portrays, generating an effective illusion including maintaining a history, managing dialog, generating posture and expressions and tracking appropriate affect is a large and significant software engineering project.

Barry Silverman at the University of Pennsylvania is pursuing a similar effort to endow less complex agents within simulations with cultural and other individual behavioral differences (Silverman, Johns, Cornwell, & O'Brien 2006a,b). His approach uses performance *moderator* functions similar to the performance shaping functions found in Micro Saint and risk assessment. In Silverman's implementation these functions are managed by a separate PMFserv application that is polled by the simulation, in the case of Silverman et al. (2006b), Soar-bots running inside the Unreal 2 game engine. Figure 8 shows a performance moderating function based on the coping styles identified by Janis and Mann (1977), an elaboration of the Yerkes-Dodson law linking performance to arousal. Note the conceptual similarities to the PSF for a nuclear accident.

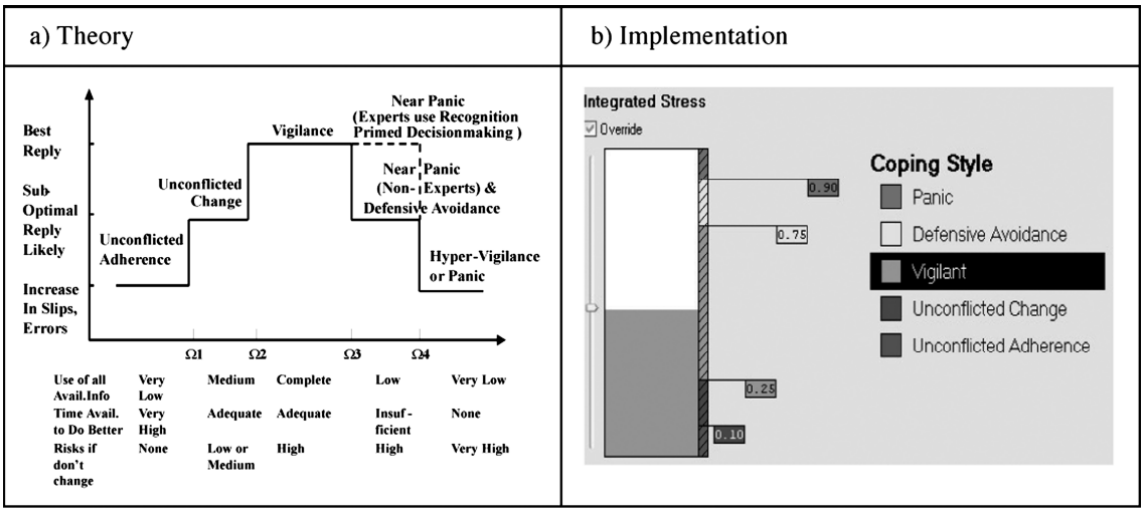

Figure 8. Performance moderating function for Janis-Man/Yerkes-Dodson reprinted from Silverman et al. (2006a).

Prediction of Team Performance

As the preceding section has shown, accurate simulation of human behavior is still in its infancy. In these examples there were generally "sweet spots" defined by granularity and types of behavior within which a given implementation did well. Outside of this range it deteriorated. ACT-R, for instance, was impressive with its predictions for low level cognitive behavior, but confronted with the complexity and time scales of MOUT tasks fell back on production rules more or less identical to those used by systems such as Micro Saint without cognitive pretenses. A key consideration in choosing agent models for Navy teams, therefore, should be the desired granularity and the behaviors and influences that need to be modeled accurately. A guide to efficiency would be to model at as coarse a level as possible while still capturing the behaviors of interest. In the case of teamwork, the behaviors and their characteristics were presented earlier. An examination of the behavior taxonomy shown in Figure 1 indicates that time could be represented loosely through the ordering of events without affecting any of the predictions. Similarly, short and long term memory do not appear to be factors. In contrast, substantial domain knowledge and the ability to classify and attribute communications and actions of others appear to be prerequisites. These requirements would argue for *weak* AI (i.e.; agents whose behavior is largely programmed and constrained rather than following general cognitive principles).

In Figure 3 (reproduced below) a conceptual model of the effects that selection might have on team behavior is suggested. In this model, individual differences, team composition, and team attributes acted to moderate the behavior of a normative team model. Team member roles, goals, and interactions must be fairly precisely defined for such a model to exist. Fortunately this is often the case for military tasks of interest.

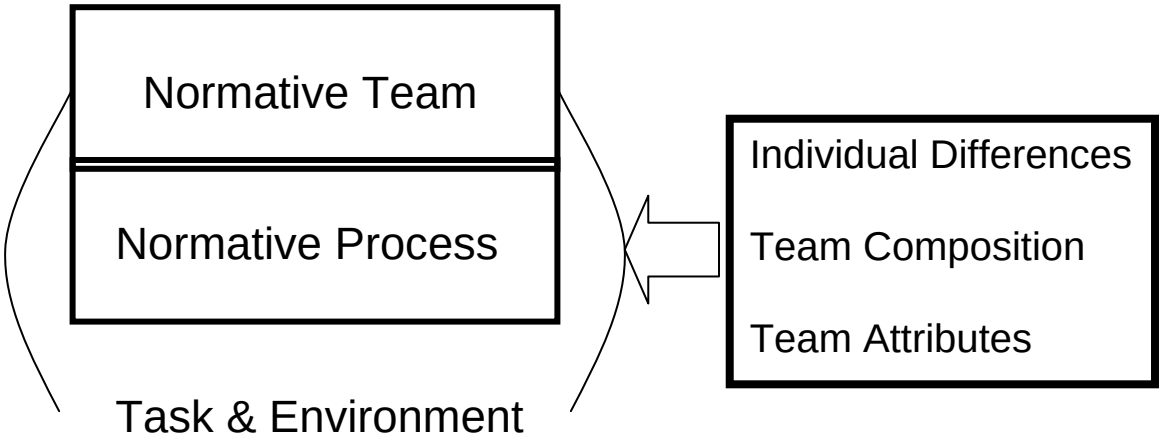

Figure 3. Teamwork model reproduced.

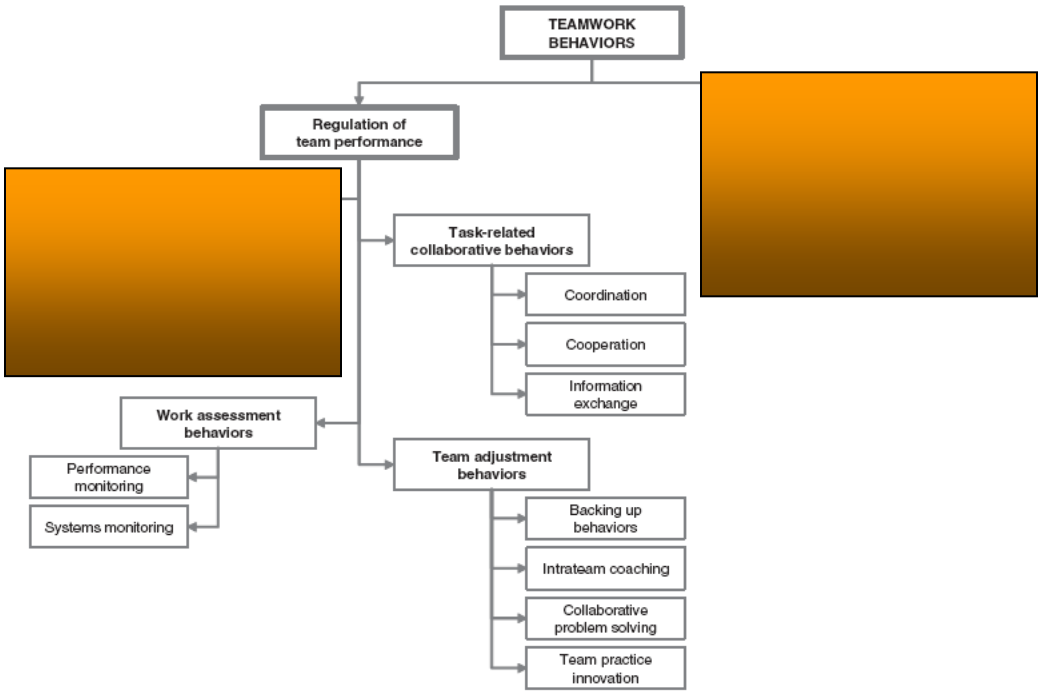

Figure 9. Rousseau's taxonomy with excluded processes in shaded areas.

Recommendations

As discussed earlier, this conceptual model is a variant of the performance shaping function approach. It requires the ability to specify at the agent (individual differences) or team (team attributes) levels the effects of the moderators. Team composition is a special case because it arises from individual differences but is expressed in available data at the team level. How this should be dealt with would need to be considered in implementation. A number of the agent models discussed would be suitable for such a normative model if it were restricted to well-constrained procedural tasks. Figure 9 shows Rousseau's taxonomy with excluded processes in the shaded areas.

RETSINA, Machinetta, COGNET/iGEN, Soar, or Micro Saint/IPME would all be suitable for this sort of modeling. Following the announced preference for simpler simulations would reduce the list to the two task network modelers: RETSINA and Micro Saint/IPME and Machinetta with its even more restrictive TOPs and built-in (though needing modification to match human) teamwork behaviors.

If modeling were extended to include planning (the preparation of work accomplishment blocks) and team adjustment behaviors requiring problem solving and learning the list would be reduced to RETSINA and Soar. In this case, considerable validation would be needed to adjust either RETSINA's HTN planning mechanism or Soar's generalization mechanisms to reflect human planning behavior.

Challenges to the Validity of Team Models

To the extent that modeling is restricted to constrained, well-practiced tasks with well defined role responsibilities and errors are limited to random omissions or commissions the normative models should be adjustable to account for observed human performance. The normative models, however, are only intended to serve as a sort of "cloud chamber" to allow observation of the effects of the performance shaping functions on team behavior. For this to work PSFs must be tightly parameterized both in their isolated effects and in their interactions. A glance at the studies reviewed earlier will reveal that the researchers are far from this goal. While the direction of the effect of a measurable variable such as team cohesion is well supported, precisely how much it should enhance or degrade the output of an executing simulation is not known. Even well established psychological laws prove difficult to quantify. The ad hoc characterization of PSFs for a nuclear accident or Janis-Mann coping styles are typical of such attempts.

In the absence of clear quantitative data to define PSFs and determine their parameters an alternative may be to use team simulations as an experimental testbed for examining the sensitivity of team performance at the extremes. The team simulation could be treated as a hypothesis generator for subsequent confirmation/disconfirmation by real data. Proceeding in such a way it might over time be possible to develop confidence in the normative and PSF models. Since this would require affirming the micro model on the basis of macro observations selecting a parsimonious (i.e., task network) normative model and restricting PSFs to a small number with pronounced effects would be necessary. Since a graphical representation of events and user interaction would be unnecessary for agent-only simulations, relatively lightweight, fast running simulations could be constructed providing a simple or modular agent architecture is chosen. This would allow generation of large test sets that systematically cross psf's to help adjust the models to observed interactions between psf's.

Team Selection

Generally, teamwork consists of two key issues: the first is team selection which is to select the correct team members from a candidate pool, and the second is task assignment which is to assign the team members to the given duties. These two issues are tightly connected and shall be addressed in alignment with the objective to optimize the team performance. This section discusses the conventional optimization approaches on team selection and assignment. Optimization methodologies refer to the mechanisms solving problems in which one seeks to minimize or maximize an objective function by optimally choosing the values of the decision variables within an allowed set. Since a goal of a team assignment is to optimize the performance of the formed team, optimization methodologies have been widely applied by researchers and practitioners.

Conventional Optimization Approaches

Team assignment can be either static or dynamic. Static team assignment is that once the team is formed, neither the team members nor their duties will change, in contrast, dynamic team composition varies in time (i.e., new team members may join and some of the existing team member may leave); their duties may also change along with the time. The representative work on static team assignment is the "Assignment Problem" (AP) studied in Operations Research, in which a mathematical program determines the optimal assignment of the agents to a given set of tasks to either maximize the total payoff or minimize the total cost. This problem is initiated by Kuhn's seminal work in 1955 (Kuhn 1955). The mathematical model for the classic assignment problem can be given as:

$$\text{Minimize} \quad \sum_{i-1}^{n} \sum_{j-1}^{n} c_{ij} x_{ij}$$

$$\text{Subject to} \quad \sum_{i-1}^{n} x_{ij} = 1 \quad j = 1,\ldots,n$$

$$\sum_{j-1}^{n} x_{ij} = 1 \quad i = 1,\ldots,n$$

$$x_{ij} = 0 \text{ or } 1$$

where $x_{ij} = 1$ if agent i is assigned to task j, 0 if not, and c_{ij} = the cost of assigning agent i to task j. The first set of constraints ensures that every task is assigned to only one agent and the second set of constraints ensures that every agent is assigned to a task. The basic mathematical structure of the problem makes the constraint that x_{ij} be binary unnecessary since there will automatically be an optimal linear programming solution in which every x_{ij} is either 0 or 1. This classic assignment problem is mathematically identical to the *weighted bipartite matching* problem from graph theory and thus results from that problem formulation have been used in constructing efficient solution procedures for the classic assignment problem.

After Kuhn's seminal work, there is a stream of research that extends the classical assignment problem by considering: agent qualification (Caron, Hansen, & Jaumard, 1999) where an agent may only be qualified for a subset of tasks, partial agent and task matching (Dell'Amico & Martello, 1997) where only a subset of given tasks need to be assigned and only a subset of the agents can be deployed, bottleneck assignment problem (Ford and Fulkerson, 1966) in which the problem is to minimize the maximum cost of assigning the tasks, the semi-assignment problem (Kennington & Wang, 1992) where some tasks that need to be assigned are identical; etc.

This stream of work on extensions of the classical assignment problem has assumed that each agent may only take one task. This might not be true in practice. Therefore, researchers have also studied generalized assignment problems (GAP). These models assume that each task will be assigned to one agent, but it allows for the possibility that an agent may be assigned more than one task, while recognizing how much of an agent's capacity each task would use. Thus, the GAP is an example of a one-to-many assignment problem that recognizes capacity limits. Recognizing that a task may use only part of an agent's capacity (GAP) rather than all of it (AP), leads to the following model:

$$\sum_{i-1}^{m} a_{ij} x_{ij} \leq b_i$$

$$\sum_{i-1}^{m} x_{ij} \leq b_i \quad j = 1,\ldots,n$$

$$\sum_{j-1}^{n} a_{ij} \leq b_i \quad i = 1,\ldots,m$$

$$x_{ij} = 0 \text{ or } 1$$

where x_{ij} = 1 if agent i is assigned to task j, 0 if not, c_{ij} = the cost of assigning agent i to task j, a_{ij} is the amount of agent i's capacity used if that agent is assigned to task j, and b_i is the available capacity of agent i. The first set of constraints ensures that every task is assigned to only one agent and the second set of constraints ensures that the set of tasks assigned to an agent do not exceed its capacity.

With the more realistic characteristics, GAP has wider applications. In particular, Garrett, Dasgupta, Silva, Vannucci, and Simien (2005) model the Navy Sailor assignment problem by the GAP model and design evolutionary algorithm solving techniques that provide efficient solutions. Similarly, Holder (2005) models Navy personnel job assignment while additionally considering Sailor satisfaction, and designs traditional optimization solving techniques.

The models discussed so far are all static models where there are no stochastic factors and the models do not consider future amendments either from the tasks' side or agents' side. For instance, in Garrett et al. (2005), the authors assume that the jobs that Sailors are assigned to are deterministic and there would not be new tasks appearing or modifications on the old tasks; similarly, Sailors also will not change things, such as their capabilities or characteristics. Therefore, a more realistic extension of the above models is to consider the uncertainties and future variations. Similar problems widely exist in practice, such as call center scheduling problems where tasks are arriving stochastically and agents may join and leave the workforce. Traditional optimization mechanisms to address this type of problems are dynamic programming (DP) and scheduling. Generally, a DP assignment model assumes that there are multiple periods in which decisions need to be made on task and agent assignment; the agents once assigned to some tasks may be occupied for some time length (e.g., they will become free again after they finish their current tasks); the future modifications follow some stochastic pattern (e.g., stochastic process); and the goal is to optimize the aggregated performance in the whole time horizon. There is extensive literature on these problems.

Mehrotra and Fama (2003) provide an extensive tutorial on call center staffing, scheduling and traditional simulation techniques. Similarly, Ernst, Jiang, Krishnamoorthy, and Sier (2004) provide a review on staff scheduling and rostering.

Advantages and Disadvantages

Optimization methodologies are rigorous with systematical proofs and precise presentations. Optimization methodologies rely on rigorous optimization theories that capture each of the considered factors by mathematical representation. With an optimization model, one can find the accurate solution with proofs to the problem and present the solution in a concrete way. Furthermore, the results of the model usually can be easily understood and followed with the mathematical presentations.

This strength of optimization methodologies also comes with limitations. To apply those models, one must be able to model the problem characteristics using mathematical representations. However, there are many factors in practice that are difficult to model in mathematics, such as Sailors' personalities, their satisfactions with the tasks, the team cohesion, and the uncertainties in Navy task execution. Therefore, this means that to follow the conventional optimization theories, one has to compromise many factors that are important in a teamwork assignment. The second limitation of the conventional optimization methodologies is that they are also constrained by the computation complexity. To solve a large size GAP or a DP program is extremely computationally expensive. Usually, exact solutions are not computationally tractable to obtain. In such cases, one has to apply heuristics that sacrifice the accuracy of the solution. Finally, the conventional optimization methodologies are all centralized programs. In other words, in those models, there is a central planner who comes up with the schedule to deploy the team members and the team members do not have any decision power but follow the assigned duties. This might not be always true in practice, particularly when one deals with people rather than machines, or in domains where such a powerful and capable central planner does not exist.

Linkage to TESTOR

Optimization methodologies, however, can still be appropriately applied in the Navy teamwork if the obstacles can be solved. In the Navy teamwork problem, one can divide the factors (or coefficients) that impact the teamwork performance into two groups: hard factors and soft factors. The hard factors refer to those that can be directly mathematically modeled, such as the number of tasks, the quantity of resources that are needed, and the monetary payoff that can be realized from finishing the tasks. The soft factors are those that cannot be modeled directly in mathematics. Those factors can include the ones discussed above, such as agent personalities, teamwork skills, team cohesions, etc. To cope with the soft factors, team simulation and psychological theories can be applied. For instance, agent-based team simulation combined with psychological theories can be applied to characterize the impacts of agent personalities, teamwork skills and team cohesion on task performance. Generally, with the simulation tool, one can assign different types of agents (with particular parameters) to some particular tasks and then summarize the realized performance. Next, checking with statistical

observations (history data), one can find which set of parameters are realistic to model the impacts of personalities, teamwork skills and team cohesion. Finally, those parameters can approximately represent the impacts of those soft factors in reality.

With all the key coefficients being characterized, an optimization model on team assignment can be developed, as either a static version for closely static problems or as a dynamic version for stochastic problems with timing consideration. In particular, to deal with the stochastic factors and computational complexity, one can decompose the original optimization problem into sub-problems that are tractable, and apply agent-based simulation to approximate the whole solution for the original problem. Moreover, based on multi-agent simulation systems, decentralized factors also can be captured by modeling the agents as autonomous decision makers, and one can simulate the performance of the team with the solution of the task assignments obtained from the optimization model; by such close-loop checking and amendment on the models, one can find the final satisfactory solution of the problem.

Individual Diagnostic Assessment of Teamwork

Surprisingly, diagnostic assessment of Sailor teamwork where a Sailor interacts with a team of agents may be easier to achieve than prediction of team performance. This occurs because, as reviewed earlier, there are well-developed criteria for assessing the quality of teamwork. Unlike all-agent team simulations which could be run without extensive computation in faster than real time, hybrid simulations involving humans and agents must provide user interfaces and present events in a compelling and realistic way. Such an application would require identifying a range of situations and scenarios that could elicit the types of teamwork behaviors to be assessed. Due to motivational factors these test scenarios should draw on skills and domain knowledge the Sailor already possesses and have sufficient realism to induce stress or other mental states of interest. Agents could be programmed to interact adaptively to provide opportunities for observing human responses such as monitoring or backing up behaviors. It may be advisable to restrict evaluation to the particular types of team organization or tasks that are the focus of interest. In addition to assessing teamwork behaviors of the testee directly, comparisons could be made between team performance for the testee's team and that of a reference team consisting of all agents or a team with a high scoring human. This section presents two approaches to assessing teamwork. The first requires understanding the role, context, and actions required of a team member and assessing these aspects of performance. The second approach avoids understanding task or teamwork demands and instead assesses performance by judging overall similarity to a reference.

Teammate Turing Test

First, the critical feature of any such system would be the ability of the agents to supply realistic enough behavior and interactions to elicit teamwork behaviors for measurement. This requires that agents must be able to perform their taskwork in a

correct and credible manner, communicate realistically with the Sailor, and comprehend both behaviors and communications from the Sailor. Difficulties in maintaining common ground so that agents and Sailor hold similar views of the state of the world will vary greatly depending on the types of tasks being simulated. Situation displays such as interactive maps provide an excellent basis for maintaining common ground because the map is available to both human and agents and human actions such as selection of objects or locations are unambiguous. Menu selections and toolbars are easy to interpret for the same reasons. Textual interfaces such as chat programs, now widely used in some military contexts, can also provide a good interaction medium provided a communications protocol and restricted vocabulary are used. (Restricted vocabulary and adherence to communications protocols, incidentally, were some of the characteristics that Prince & Salas [1993] found distinguished effective teams.)

A second challenge affecting the difficulty of simulating agent teammates involves the degree of constraint provided by the task. For highly constrained tasks or procedurally driven checklists, both errors of omission and commission are more easily identifiable. Because role following dictates where an action or communication should occur as well as its general form, a program can check responses against a lattice that orders the tasks to enforce necessary orderings and use string matching to assess content. Table 2 shows criteria that might easily be assessed automatically from the proposed criteria for assessing the quality of group processes presented in Figure 2.

Table 2
Teamwork criteria that might be assessed automatically

Sensing

Attempts to obtain information are relevant to mission, task, or problem

Communicating Information

- Timeliness appropriate to unit requirements
- Correct choice of recipients; everyone who needs information receives it
- Whether message should have been communicated

Decision making

Appropriateness: Timing is appropriate in view of the situation, conditions, and intended effects. Choice of target of the action is appropriate.

Communicating implementation

Transmission to appropriate recipients

Coping actions

Timeliness of the action in view of both operational circumstances and the decision or order from which the action derives

Feedback

Timeliness of the feedback decision and action

As noted, syntactically-based judgments involving timing, choice of recipient, accesses to obtain information, or relaying of information might be automated with relative ease. Semantic judgments requiring assessment of accuracy, adequacy, or appropriateness would be substantially more difficulty. This would be particularly true for spoken communication where recipient and time would remain easy to measure but semantic judgments would be made more difficult both by errors in speech recognition and the tendency for verbal responses to be less restrained.

To illustrate these distinctions we will compare two team tasks previously used in hybrid human-agent team experiments, Tandem (Sycara & Lewis, 2002) and Moksaf (Sycara & Lewis, 2004).

TANDEM is a moderate fidelity simulation of a target identification task, jointly developed at the Naval Air Warfare Center-Training Systems Division and the University of Central Florida. TANDEM simulates cognitive characteristics of tasks performed in the command information center (CIC) of an Aegis missile cruiser. Figure 10 shows a typical TANDEM display. Information about the hooked target (highlighted asterisk) is obtained from the pull-down menus *A,B,* and C.

The cognitive aspects of the Aegis command and control tasks which are captured include time stress, memory loading, data aggregation for decision making, and the need to rely on and cooperate with other team members to successfully perform the task. In performing the task subjects must identify and take action on a large number of targets (high workload). The simulation consists of three networked personal computers each providing access through menus to five parameters relative to a “hooked” target. Subjects must communicate among themselves to exchange parameter values in order to classify the target.

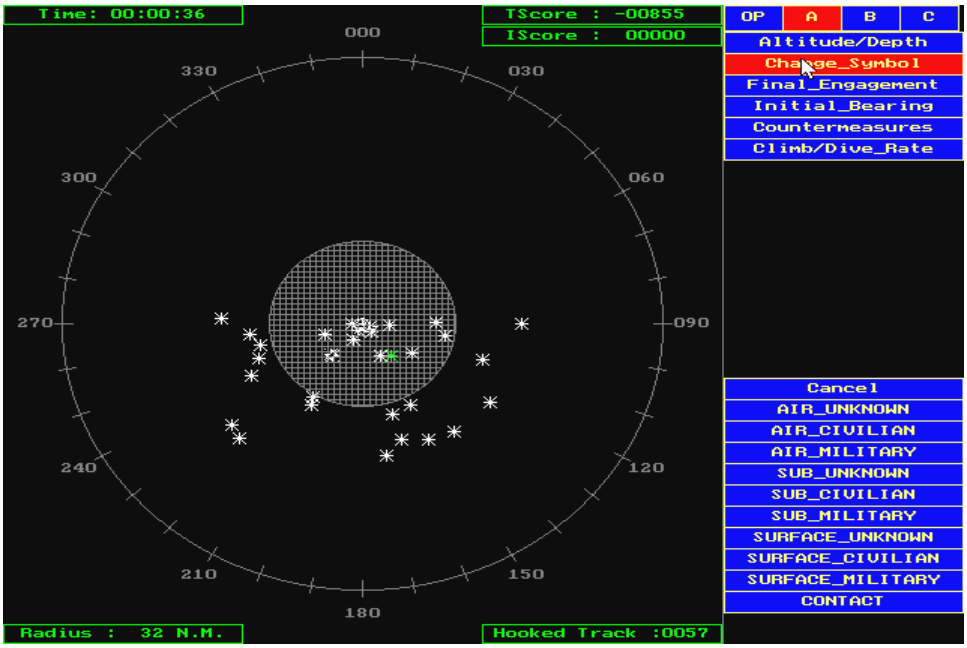

Figure 10. Tandem display.

MokSAF (Figure 11) is a simplified version of a virtual battlefield simulation called ModSAF (modular semi-automated forces). *MokSAF* allows three commanders to interact with one another to plan routes over a particular terrain. Each commander is tasked with planning a route from a starting point to a rendezvous point by a certain time. The individual commanders must then evaluate their plans from a team perspective and iteratively modify their plans until an acceptable team solution that brings the proper composition of forces with adequate supplies to the rendezvous point is developed.

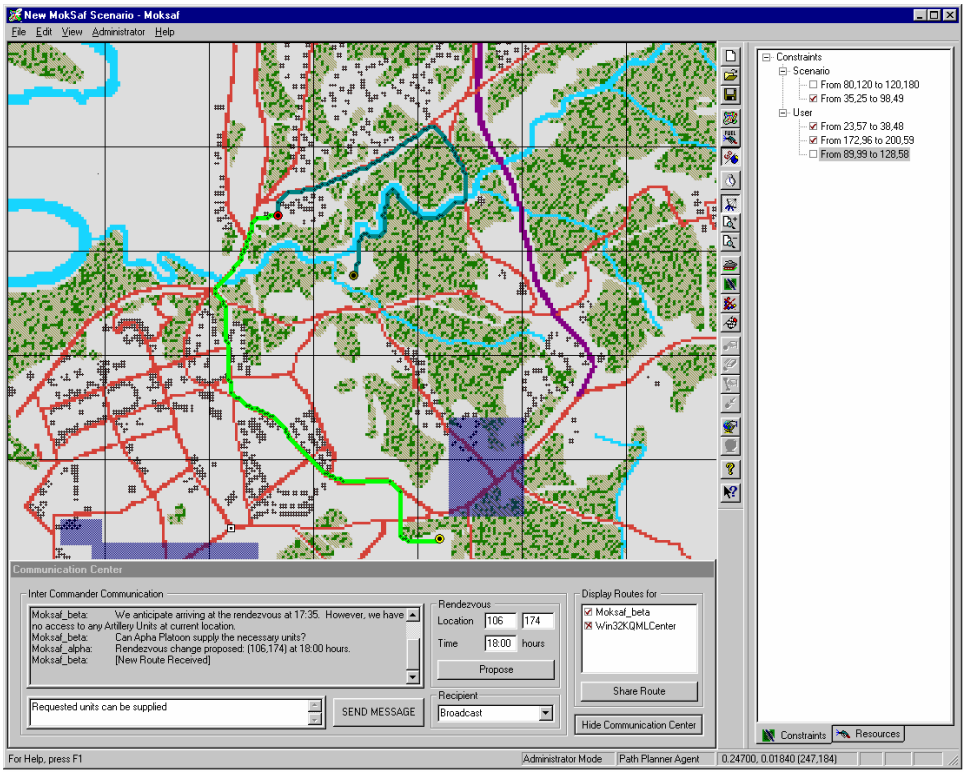

Figure 11. MokSAF display.

Table 3 contrasts the two tasks. While in TANDEM it is easy to determine what information is needed by which player and whether it has been exchanged there is no similar template for judging performance in Moksaf.

Table 3
Comparison of highly constrained and loosely constrained simulations

Tandem—highly constrained with easily classifiable behaviors	Moksaf—loosely constrained with natural language communications and problem solving
Tandem radar task with communications through chat • **Constrained communications:** agent can extract or communicate parameter name & value with little uncertainty. Selection of targets on screen is unambiguous.	Mission planning task with natural language interface • **Unconstrained communications:** agent cannot easily interpret communications because they are not tightly restricted by context.
Common Ground: simulation state and knowledge of what testee has viewed allow agent to judge human state, choose an appropriate response and judge the appropriateness of the human's response in turn	**Lack of Common Ground:** Because there is insufficient context to interpret mouse movements, clicks, utterances, etc. it is more complex to program agent to respond as a teammate

As with the all-agent team simulation the choice of agent architectures will be dependent on the required capabilities. Agents using task networks would again suffice for simulating teammates and assessing performance at highly constrained tasks. If agents are required to simulate human teammates at less structured tasks requiring problem solving, model tracing to infer states of the human testee, and relatively unconstrained communications the problem becomes much more difficult and would require a large scale development effort.

Similarity-based Assessment

Recent work applying latent semantic analysis (LSA) offers some hope that the quality of teamwork behavior might be identified from voice communications without requiring natural language understanding. Foltz, Martin, Abdelali, Rosenstein, and Oberbreckling (2006) report a correlation $r = .76$ ($p < .01$) predicting performance scores based on similarities in dialog and patterns of communications among teams performing a UAV control task. Of more interest for diagnosing teamwork behaviors, they report success in tagging communications finding a Kappa equal to .48 for agreement with human raters. Analysis of an older data set augmenting the LSA measure with additional natural language measures and selecting the best subset led to correlations of between .45–.78 with subject matter expert (SME) ratings for 16 teamwork behaviors. It is too early to predict whether such hand-tuned methods could be adapted to automated online analysis or whether they would be able to perform as well with data varying by only a single testee. The possibilities, however, are intriguing especially for tagging which could provide a basis for both diagnosis and feedback.

Simulation Test Environment

Selection of the simulation test environment(s) should depend on the aspects of teamwork and work context to be assessed. Table 4 contrasts the two general classes of simulations that might be appropriate.

Table 4
Simulation environments

Discrete Event	Real time (continuous)
• Adaptable to a wide variety of tasks • Suitable for textual or graphical interfaces • Easy to log and program interactions • Not suited for psycho-motor tasks • Does not provide immersion or presence	• Physical fidelity (e.g. flight simulation, assembly & repair, etc.) • Appropriate for stressful, reactive tasks • Requires 3D graphical interface for best effect • May generate voluminous logs • May provide immersion or presence • Is scalable to HMD/Cave environments

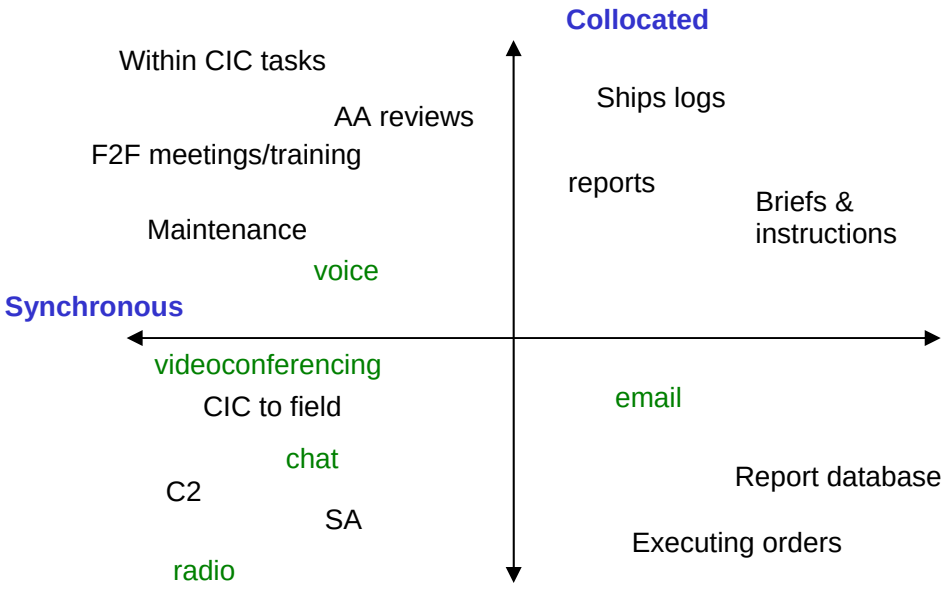

Figure 12. Time and space distinctions commonly made in CSCW.

Figure 12 provides a commonly used computer-supported cooperative work (CSCW) categorization of group tasks in terms of participant location and timing of interaction. For all of this figure except the upper left quadrant, humans could be replaced by agents without change to the appearance of the task. These cases in which participants are separated by space or time also generally involve cooperative tasks which are mediated electronically obviating the need for model physics, facial expressions, or other continuous events. This makes discrete event simulation a logical choice for such tasks. Although discrete event simulations are simple enough to develop one specifically for this purpose there are many available that could be adapted. Distributed Dynamic Decision making (DDD) developed by Aptima shown in Figure 13, for example, is a configurable simulation providing a map-based display and suitable for simulating a variety of C3I tasks. Discrete event simulations we have developed include MokSAF (Sycara & Lewis, 2004) for route planning, Morse (Sycara, Scerri, Giampapa, Srinivas, & Lewis, 2005) for NASA range operations, and Sanjaya (Scerri, Owens, Yu, & Sycara, 2007) for UAV and ground operations.

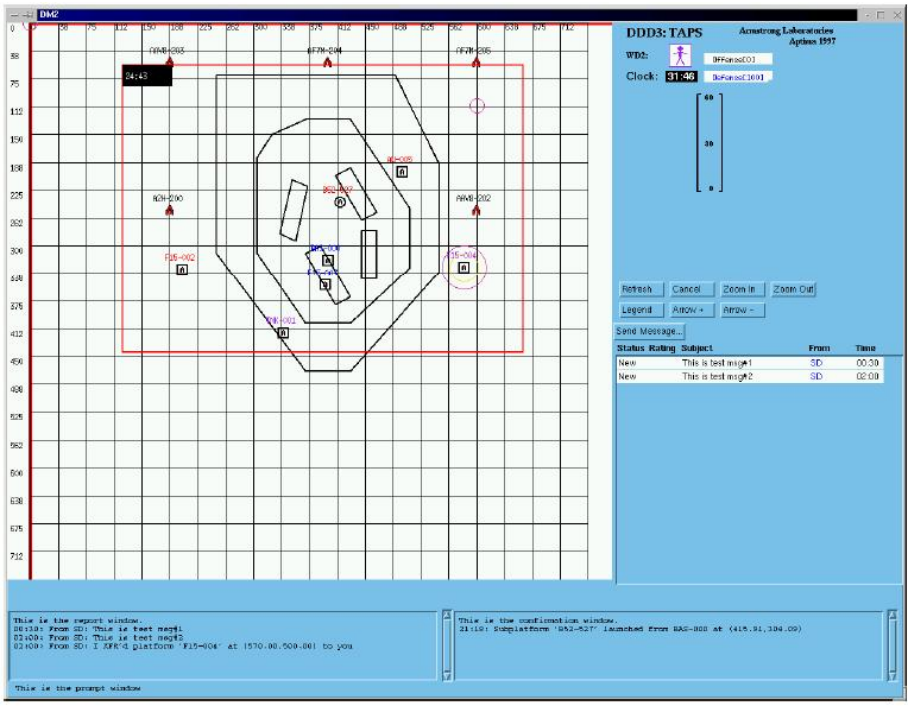

Figure 13. AWACS display simulated in DDD.

If tasks need to involve face-to-face interactions, or require “out the windshield” views to induce stress or temporal demands, a continuous simulation would be needed. Unlike discrete time simulations which are fairly simple to construct and integrate with other applications, continuous simulations require extensive software and are difficult to develop and instrument. If a continuous simulation is needed we strongly recommend adapting an existing game engine for this purpose. There are a variety of available engines ranging from the opensource Delta3D (www.delta3d.org) developed at the Naval Postgraduate School to extremely expensive proprietary game engines such as

Epic Games Unreal 3 Engine (http://www.unrealtechnology.com/licensing.php). There are also so called *strategy* games such as the open source Global Conflict Blue (http://gcblue.com/) that mix aspects of both continuous and discrete event simulation.

Conclusions and Recommendations

The scope of agent development effort will primarily depend upon a number of choices:

First, the degree of sophistication in agent reasoning. Some of the choices are:

- Normative procedural models and performance shaping factors for them
- Planning, problem solving
- Team adjustment behaviors
- Team maintenance (meta control over repeated episodes)

Second, the intended functionality of the simulation. Some of the choices are:

- Prediction of team performance
- Team selection
- Individual diagnostic assessment of teamwork
- Assessment of teamwork for human teams

Third, the type of simulation (discrete event vs continuous) and the human interface to the simulation.

The primary determinant of level of effort will be the choice between a normative procedural agent model and one capable of less constrained behavior including problem solving and learning. This effort would involve not only construction and programming of agents but also calibration and validation of agent behaviors. We anticipate that calibration and validation would be substantially more expensive than the initial programming particularly for more sophisticated/less constrained agents. The difference in effort between procedural models and models that include problem solving and learning is because normative procedural models can be calibrated against variations in human performance associated with performance shaping factors and their interactions. For less constrained behaviors the range of possibilities becomes so great that new sampling and estimation methods would need to be developed for calibrating and testing agent models. Even then, with so many degrees of freedom these models are likely to overfit the data making performance prediction difficult.

The choice between discrete event and continuous simulation types should have a smaller impact on level of effort although discrete event simulations are easier to design, program, and interface with agents. Finally, if the simulation is used to assess human performance, a human-computer interface and methods for assessing performance will be needed. This would add additional costs to the project.

Table 5 shows the relative levels of effort projected for the alternatives that arise when considering different combinations of the three types of considerations, namely agent reasoning, intended functionality of the simulation and simulation environment. The grayed out cells indicate alternatives unlikely to contribute to TESTOR'S objectives. For example, running agents in an only-agent simulation in a continuous simulation environment is not advisable since the requisite technology (e.g., imbuing agents with human perceptual capabilities, sophisticated collision avoidance, and path planning, etc.) is not routinely available; hence this type of development would be very expensive without giving proportional benefit. On the other hand, discrete event simulation is a reasonable alternative for agent-only simulations since (a) the development methodology is available, and (b) the needed data could be collected efficiently by running the agents in faster than real time. Conversely, in time stressed tasks for which humans need continuous simulation, tasks are predominately constrained and procedural making sophisticated agent teammates unnecessary.

Table 5
Projected levels of effort

Agent Sophistication	Discrete Event		Continuous	
	agent only	agent + humans	agent only	agent + humans
Normative procedural	Alternative-1 Low	Alternative-2 Moderately Low		Alternative-3 Moderate
Interleaved planning & execution	Alternative-4 Moderately High	Alternative-5 High		

Alternative-1. *Agent Only* Prediction of Team Performance with DE Simulation and Procedural Tasks

The primary effort involved in this alternative would be collecting data and validating models for the effects and interactions of performance shaping factors in procedural tasks. While some data (reviewed previously) on the effects of individual factors are available, very little is known about their interactions. How, for example, would the distribution of mental ability, extraversion, team cohesion, and task skills interact to influence team performance? To construct a computational model, these contributions would need to be specified precisely. This is not available from the current literature and would require estimates from subject matter experts, new survey items, focus groups, or other sources. Once constructed the models would require validation. This would need to be repeated on a task by task basis until a representative sample (~10+) of tasks has been modeled.

Alternatively, focusing on a target task or group of tasks of particular interest to the Navy might allow more rapid and accurate modeling but only for restricted types of teams and tasks.

Effort estimates for Alternative-1.

Data collection and generation for modeling 2 man years/task @ 10 tasks:	20 man years
Model validation 1 man year/task @ 10 tasks:	10 man years
Model construction 1/10 man year/task @ 10 tasks:	1 man year
	31 man years

Alternative-2. Agent Model for *Prediction and Human Participation* for Assessment Using DE Simulation and Procedural Tasks

Development costs for Alternative-2 include all data collection, modeling and validation costs for Alternative-1. In order to interact with humans and assess human teamwork there are additional requirements for the development of (a) a human-computer interface, and (b) methodologies and software for assessing hybrid teamwork performance. Unlike an agent-only simulation which only needs to support message passing and events, a simulation interacting with humans needs to provide a human-agent interface. The interface must display graphical and other information to the human and interpret human inputs to the system and agent teammates. The effort involved will depend on the character of this interaction. If a "shared" graphical interface such as a radar or map display is used and communication comes primarily through interacting with this display by doing things such as selecting or classifying targets, the effort should be moderate. Designing and implementing displays of this sort for discrete event simulation is relatively easy. An existing simulation such as DDD could be adapted or a new simulation developed in-house for this purpose. The keys to limiting development effort are (a) making human inputs intelligible to the agents by interacting through a shared display and (b) limiting the richness of human-agent interaction by restricting communications to predictable referents on the screen. This allows the system to match human behaviors against those expected from a team member performing appropriate teamwork. An alternative or parallel assessment of teamwork might be provided by automated communication analysis. Although automated communications analysis provides a less accurate assessment of teamwork than direct measurement of agreement with appropriate actions, it can be used in situations where a reference model has not been developed. Using communication analysis would add the additional costs of programming agents to generate appropriate textual or verbal communications and require validation of the measures for use in teams incorporating synthetic teammates. We estimate that incorporating automated communications analysis would require 5-10 man years in addition to the effort estimates shown below.

Effort estimates for Alternative-2:

Alternative-1	31 man years
Interface and simulation development:	3 man years
Teamwork scoring & assessment (10 tasks):	2 man years
Agent interpretation of human input and communication generation:	3 man years
	39 man years

Alternative-3. Agent Model for *Prediction and Human Participation* for Assessment Using *Continuous Simulation* and Procedural Tasks

Development costs for Alternative-3 include all data collection, modeling and validation costs for Alternative-1. Third party tools such as a game engine or simulation environment such as Olive (http://www.Forterrainc.com) would be needed to develop effective 3D continuous simulations. Interfacing agents with continuous environments requires significantly greater effort than for discrete event simulations as indicated in the estimates. Provided that most interaction is via the simulated environment and communications are restricted, these costs should remain similar to those of Alternative-2. Assessing teamwork using automated communications analysis could be appropriate for Alternative-3 and we would again predict 5–10 man years of effort in addition to the effort estimates shown below.

Effort estimates for Alternative-3:

Alternative-1	31 man years
Interface and simulation development:	10 man years
Teamwork scoring & assessment (10 tasks):	2 man years
Agent interpretation of human input and communication generation:	5 man years
	48 man years

Alternative-4. *Agent-Only Prediction* of Team Performance with DE Simulation and *Interleaved Planning and Execution*

Procedural tasks are relatively easy to model and validate because they prescribe particular actions under particular conditions. Determining the effect of a PSF requires only determining the change in an action or its probability under different levels of the PSF. Even some forms of archival or retrospective report data might be used although dynamic aspects of team performance could be obscured.

Where behavior is not fixed but may vary widely while remaining appropriate, as in Intelligence Preparation of the Battlespace, it becomes much more difficult to model. This is not because planning algorithms are so difficult to implement but because it is very difficult to verify that a planning program will make the same choices and errors as the human(s) being modeled. Unlike a procedural model which could be validated against multiple repetitions of the same task by different teams, a planning/problem solving model would need to be validated against a sample of problems from the population of possible problems. Each problem of this sample would in turn require its own repetitions by human teams for validation. Plausible models could be programmed and run with moderate effort, however, the validity of their predictions would not be known. Whatever the approach to this alternative it would probably be advisable to pick a relatively restricted team and problem/task type.

Effort estimates (validated models) for Alternative-4:

Data collection and generation:	25 man years
Model validation:	30 man years
Model construction:	10 man year
	65 man years

Alternative-5. Agent Model for *Prediction and Human Participation* for Assessment with DE Simulation and *Interleaved Planning and Execution*

Development costs for Alternative-5 include all data collection, modeling and validation costs for Alternative-4. The data collection and modeling needed to validate agent models would provide a ready reference for assessing human trainees. The costs of programming agents to generate appropriate textual or verbal communications; however, would be substantial and require extensive testing because of the lack of constraints on communications. These difficulties would be accentuated if automated communications analyses were contemplated.

Effort estimates (validated models) for Alternative-5:

Data collection and generation:	25 man years
Model validation:	30 man years
Model construction:	10 man year
Teamwork scoring & assessment (10 tasks):	5 man years
Agent interpretation of human input and communication generation:	10 man years
	80 man years

These estimated levels of effort are very rough approximations and intended to give a sense of the relative difficulties. Many of the development activities could be performed in parallel. We advise adopting an incremental approach to development due to the innovative nature of the proposed systems. We foresee model validation as the greatest challenge and believe that developing a pilot prototype would be advisable. This prototype could be used to help determine the forms of data needed and testing required to attain the desired levels of prediction from team models.

We believe that the pilot effort should start by selecting a procedural team task of interest to the Navy (from Alternative-1) for which substantial data on process as well as outcomes already exist. Although ultimately team models are to be developed and tested using forms of data most readily available, we believe that it is crucial to start with a task that can be simulated in the laboratory. This would allow developers to test hypotheses about mechanisms as well as outcomes in order to develop an accurate model of the task and performance shaping factors. This reference model could be used in turn to help identify data requirements and expected quality of prediction for models built using other types of data. Results from this pilot should provide more accurate assessments of the costs and expected ROI for full implementation of one or more of the alternatives.

References

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

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44, 1–26.

Barrick MR, Stewart GL, Neubert MJ, & Mount MK. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology*, 83, 377–391.

Barry, B., & Stewart, G. L. (1997). Composition, process, and performance in self-managed groups: The role of personality. *Journal of Applied Psychology*, 82, 62–78.

Beal, D.J., Cohen, R.R., Burke, M.J., & McLendon, C.L. (2003). Cohesion and performance in groups: A meta-analytic clarification of construct relations. *Journal of Applied Psychology*, 88, 989–1004.

Best, B., Lebiere, C., & Scarpinatto, C. (2002). A model of synthetic opponents in MOUT training simulations using the ACT-R cognitive architecture. In *Proceedings of the Eleventh Conference on Computer Generated Forces and Behavior Representation*. Orlando, FL.

Caron, G., Hansen, P., & Jaumard, B. (1999). The assignment problem with seniority and job priority constraints, *Operations Research* 47(3), pp. 449–454.

Carr, J.Z., Schmidt, A.M., Ford, J.K., & DeShon, R.P. (2003). Climate perceptions matter: A meta-analytic path analysis relating molar climate, cognitive and affective states, and individual level work outcomes. *Journal of Applied Psychology*, 88, 605–619.

Chen, G., Donahue, L.M., & Klimoski, R.J. (2004). Training undergraduates to work in organizational teams. *Academy of Management Learning and Education*, 3, 27–40.

Cognet (2008). http://www.chisystems.com/ accessed January 23, 2008.

Cohen, P., & Levesque, H. (1990). Intention is choice with commitment. *Artificial Intelligence*, 42, 1990.

Cooke, N., Salas, E., Kiekel, P., & Bell, B. (2004). Advances in measuring team cognition, In E. Salas and S. Fiore (Eds.) *Team Cognition: Understanding the Factors that Drive Process and Performance*, Washington, DC: American Psychological Association, 83-106.

Dell'Amico, M. & Martello, S. (1997). The k-cardinality assignment problem, *Discrete Applied Mathematics* 76(1–3), pp. 103–121.

Deutsch, S., Pew, R., Tenney, Y., Diller, D., Godfrey, K., Spector, S., Benyo, B., & Date, S. (2004). *Agent-based modeling and behavior representation (AMBR)*, AFRL-HE-WP-TR-2004-0191.

Devine, D. (2002). A review and integration of classification systems relevant to teams in organizations, *Group Dynamics: Theory, Research, and Practice*, 6(4), 291-310.

Driskell, J., Salas, E., & Hogan, R. (1987). *A taxonomy for composing effective naval teams*, NAVTRASYSCEN TR87-002.

Ellis, A.P.J., Bell, B.S., Ployhart, R.E., Hollenbeck, J.R., & Ilgen, D.R. (2005). An evaluation of generic teamwork skills training with action teams: Effects on cognitive and skill-based outcomes. *Personnel Psychology*, 58, 641–672.

Ernst, A., Jiang, H., Krishnamoorthy, M., & Sier, D. (2004). Staff scheduling and rostering: A review of applications, methods and models, *European Journal of Operational Research* 153, 3-27.

Foltz, P., Martin, M., Abdelali, A., Rosenstein, M., & Oberbreckling, R. (2006). Automated team discourse modeling: Test of performance and generalization, *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Ford Jr., L. & Fulkerson, D. (1966). *Flows in Networks*, Princeton University Press, Princeton, NJ (Section II.6).

Garrett, D. Dasgupta, D., Silva, R., Vannucci, J., & Simien, J. (2005). Genetic Algorithms for the Sailor Assignment Problem, Proceedings of the 2005 *Genetic and Evolutionary Computation Conference*.

Griffith, J. (1997). Test of a model incorporating stress, strain, and disintegration in the cohesion-performance relation. *Journal of Applied Social Psychology*, 27, 1489-1526.

Griffith, J. and Vaitkus, M. (2000). Relating cohesion to stress, strain, disintegration, and performance: An organizing framework. Military Psychology, 11(1), 27-55.

Grosz, B. and Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, 86, 1996.

Gully, S.M., Incalcaterra, K.A., Joshi, A., & Beaubien, J.M. (2002). A meta-analysis of team-efficacy, potency, and performance: Interdependence and level of analysis as moderators of observed relationships. *Journal of Applied Psychology*, 87, 819–832.

Gully, S.M., Devine, D.J., & Whitney, D.J. (1995). A meta-analysis of cohesion and performance: Effects of levels of analysis and task interdependence. *Small Group Research*, 26, 497–520.

Hackman, J.R. (1987). The design of work teams. In J. Lorsch (Ed.), *Handbook of organizational behavior* (pp. 315–342). New York: Prentice Hall.

Hackman, J.R. (1992). Group influences on individuals in organizations. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, pp. 199–267). Palo Alto, CA: Consulting Psychologist Press.

Helmreich, R.L., & Foushee, H.C. (1993). Why crew resource management? Empirical and theoretical bases of human factors training in aviation. In E.L.Weiner, B.G. Kanki,&R.L. Helmreich (Eds.), *Cockpit resource management* (pp. 3–45). San Diego, CA: Academic Press.

Hill, G. W. (1982). Group versus individual performance: Are N + 1 heads better than one? *Psychological Bulletin*, 91, 513–539.

Hill, R., Belanich, J., Lane, H., Core, M., Dixon, M., Forbell, E., Kim, J., & Hart, J. (2006). Pedagogically structured game-based training development of the Elect Bilat simulation. 25th *Army Science Conference* (Orlando, FL, November 2006).

Hill, R., Chen, J., Gratch, J., Rosenbloom, P., & Tambe, M. (1997) Intelligent agents for the synthetic battlefield: {A} company of rotary wing aircraft, *Innovative Applications of Artificial Intelligence*.

Hofmann, D.A., & Stetzer, A. (1996). A cross-level investigation of factors influencing unsafe behaviors and accidents. *Personnel Psychology*, 49, 307–339.

Holder, A. (2005). Navy Personnel Planning and the Optimal Partition, *Operations research*, 53 (1), 77-89.

Holland, J. (1985). *Making vocational choices: A theory of vocational personalities and work environments* (2nd ed.) Englewood Cliffs, NJ: Prentice-Hall.

Hollnagel, E. (2000). Modeling the orderliness of human action, In: *Cognitive Engineering in the Aviation Domain*, Eds. S. Sarter & R. Amalberti, Lawrence Earlbaum Assoc: NJ, pp. 65-98.

Hough, L. M. (1992). The "Big-Five" personality variables—construct confusion: Description versus prediction. *Human Performance*, 5, 139–155.

Janis, I. L., & Mann, L. (1977). Decision making: *A psychological analysis of conflict, choice, and commitment*. New York: The Free Press.

Kalick, S. M. & Hamilton, T. E. (1986). The matching hypothesis reexamined. *Journal of Personality and Social Psychology*, 51(4), 673-682.

Kennington, J. & Wang, Z. (1992). A shortest augmenting path algorithm for the semi-assignment problem, *Operations Research* 40(1), pp. 178–187.

Kozlowski, S. & Ilgen, D. (2006). Enhancing the effectiveness of work groups and teams, *Psychological Science in the Public Interest*, Volume 7, Issue 3, Page 77-124, Dec 2006

Kuhn, H. (1955). The Hungarian method for the assignment problem, *Naval Research Logistics Quarterly* **2**(1&2), pp. 83–97.

Marks, M.A., Mathieu, J.E., & Zaccaro, S.J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, 26, 356–376.

McGrath, J.E. (1964). *Social psychology: A brief introduction*. NewYork: Holt, Rinehart, & Winston.

Mehrotra, V. & Fama, J. (2003). Call center simulation modeling: methods, challenges, and opportunities, *Proceedings of the 2003 Winter Simulation Conference*, 135-143.

Microsaint (2008). http://www.alionscience.com/index.cfm?fuseaction=Products.view&productid=35, accessed January 23, 2008.

Millitello, L., Kyne, M., Klein, G., Getchell, K. & Thordsen, M. (1999). A synthesized model of team performance, *International Journal of Applied Ergonomics*, 3(2), 131-158.

Newell, A. (1990). *Unified Theories of Cognition*, Cambridge, MA: Harvard University Press.

Olmstead, J. (1992). *Battle staff integration* (IDA Paper P-2560). Alexandria, VA: Institute for Defense Analysis.

Parunak, H., Savit, R., & Riolo. R. (1998). Agent-based modeling vs. Equation-based modeling: A case study and users' guide. *Proceedings of Workshop on Multi-agent systems and Agent-based Simulation (MABS'98)*, Springer.

Peeters, M., Rutte, C., van Tuijl, H., & Reyman, I. (2006). The big five personality traits and individual satisfaction with the team, *Small Group Research*, 37(2), 187-211.

Prince, C. & Salas, E. (1993). Training and research for teamwork in the military aircrew. In E. Wiener, B. Kanki, & R. Helmreich (Eds.), *Cockpit resource management*, San Diego, CA: Academic Press, 337-366.

Rao, A. S., & Georgeff. M. P. (1991). *Toward a Formal Theory of Deliberation and Its Role in the Formation of Intentions*. Technical Report, Australian Artificial Intelligence Institute, Victoria, Australia.

Reece, D. A. (2003). Movement behavior for soldier agents on a virtual battlefield. Presence: *Teleoperators and Virtual Environments* 12, 4 (Aug. 2003), 387-410.

Rosenbloom, P., Laird, J., & Newell, A. (1993). SOAR: an architecture for general intelligence, *Artificial Intelligence*, 33(1), 1-64.

Rousseau, V., Aube, C., & Savoie, A. (2006). Teamwork behaviors: A review and integration of frameworks, *Small Group Research*, 37(5), 540-570.

Salas, E., Dickinson, T.L., Converse, S.A., & Tannenbaum, S.I. (1992). Toward an understanding of team performance and training. In R.W. Swezey & E. Salas (Eds.), T*eams: Their training and performance* (pp. 3–29). Norwood, NJ: Ablex.

Scerri, P., Owens, S., Yu, B., & Sycara, K. (2007). A decentralized approach to space deconfliction, in *Proceedings of FUSION 2007*.

Scerri, P., Pynadath, D., Schurr, N., Farinelli, A., Gandhe, S., & Tambe, M., (2004). Team Oriented Programming and Proxy Agents: The Next Generation, *Proceedings of 1st international workshop on Programming Multiagent Systems*, Springer, LNAI 3067.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1, 143-186.

Schneider, B., & Bowen, D.E. (1985). Employee and customer perceptions of service in banks: Replication and extension. *Journal of Applied Psychology*, 70, 423–433.

Schneider, B., White, S.S., & Paul, M.C. (1998). Linking service climate and customer perceptions of service quality: Test of a causal model. *Journal of Applied Psychology*, 83, 150–163.

Shils, E. and Janowitz, M. (1948). Cohesion and disintegration in the Wehrmacht in World War II. *Public Opinion Quarterly*, 12, 280-315.

Shoham, Y. (1993). Agent-Oriented Programming. *Artificial Intelligence* 60(1): 51-92.

Siebold, G. (2007). The essence of military group cohesion, *Armed Forces and Society*, 33(2), 286-295.

Siegel, A.I. & Wolf, J.J.. (1967). *Man-Machine Simulation Models*. New York: John Wiley & Sons.

Silverman, B. G., Johns, M., Cornwell, J., & O'Brien, K. (2006a). Human behavior models for agents in simulators and games: Part I—Enabling science with PMFserv. *Presence: Teleoperators and Virtual Environments*, 15(2), 139–162.

Silverman, B. G., Johns, M., Cornwell, J., & O'Brien, K. (2006b). Human behavior models for agents in simulators and games: Part II—Gamebot engineering with PMFserv. *Presence: Teleoperators and Virtual Environments*, 15(2), 163–185.

Simon, H. A. (1957). Models of man: Social and rational. New York: John Wiley and Sons.

Smith, E., & Conrey, F. (2007). Mental representations as states not things: Implications for implicit and explicit measurement. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 247-264). New York: Guilford Publications.

Steiner, I.D. (1972). *Group process and productivity*. New York: Academic Press.

Stevens, M.J., & Campion, M.A. (1994). The knowledge, skill, and ability requirements for teamwork: Implications for human resource management. *Journal of Management*, 20, 503–530.

Stevens, M.J., & Campion, M.A. (1999). Staffing work teams: Development and validation of a selection test for teamwork settings. *Journal of Management*, 25, 207–228.

Stewart, G.L., & Barrick, M.R. (2004). Four lessons learned from the person-situation debate: A review and research agenda. In D.B. Smith & B. Schneider (Eds.), *Personality and organizations* (pp. 61–87). Mahwah, NJ: Erlbaum.

Stone, P. & Veloso, M. (1999). Task decomposition, dynamic role assignment, and lowbandwidth communication for real-time strategic teamwork. *Artificial Intelligence*,110, 1999.

Swain, A.D., & Guttman, H.E. (1983). *Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Application*, NUREG/CR-1278. Washington DC: U.S. Nuclear Regulatory Commission.

Sycara, K. & Lewis, M. (2002).Integrating agents into human teams. *Proceedings of the Human Factors and Ergonomics Society's 46th Annual Meeting*, Baltimore MD.

Sycara, K. & Lewis, M. (2004). Integrating intelligent agents into human teams. In E. Salas and S. Fiore (Eds.) *Team Cognition: Understanding the Factors that Drive Process and Performance*, Washington, DC: American Psychological Association, 203-232.

Sycara. K, Paolucci, M, Giampapa, J., & van Velsen, M. (2001). The RETSINA MAS infrastructure in the *Journal of Autonomous Agents and Multiagent Systems*.

Sycara, K., Scerri, P., Giampapa, J., Srinivas, S., & Lewis, M. (2005). Task allocation in teams for launch and range operations, *Proceedings of the 12th International Conference on Human Computer Interaction (HCII'05)*

Tambe, M. (1997). Towards flexible teamwork. *Journal of AI Research*, 7.

Tenney, Y. & Spector, S. (2001). Comparisons of HBR Models with Human-in-the-Loop Performance in a Simplified Air Traffic Control Simulation with and without HLA Protocols: Task Simulation, Human Data and Results, *Proceedings of the 10th Conference on Computer Generated Forces & Behavior Representation*, Norfolk, VA, 15-17 May 2001

Tziner, A., & Eden, D. (1985). Effects of crew composition on crew performance: Does the whole equal the sum of its parts? *Journal of Applied Psychology*, 70, 85–93.

Walster, E., Aronson, V., Abrahams, D., & Rottmann, L. (1966). Importance of physical attractiveness in dating behavior. *Journal of Personality and Social Psychology*, *4*, 508–516.

Williams, W. M., & Sternberg, R. J. (1988). Group intelligence: Why some groups are better than others. *Intelligence*, 12, 351–377.

Wooldridge. M., & Jennings, N. (1995). Intelligent Agents: Theory and Practice. *Knowledge Engineering Review* 10(2): 115-152.

Zachary, W., Santarelli, T., Ryder, J., Stokes, J., & Scolaro, D. (2001). Developing a multi-tasking cognitive agent using the COGNET/iGEN integrative architecture. In *Proceedings of 10th Conference on Computer Generated Forces and Behavioral Representation*, (pp. 79-90). Norfolk, VA: Simulation Interoperability Standards Organization.

Distribution

AIR UNIVERSITY LIBRARY
ARMY RESEARCH INSTITUTE LIBRARY
ARMY WAR COLLEGE LIBRARY
CENTER FOR NAVAL ANALYSES LIBRARY
DEFENSE TECHNICAL INFORMATION CENTER
HUMAN RESOURCES DIRECTORATE TECHNICAL LIBRARY
JOINT FORCES STAFF COLLEGE LIBRARY
MARINE CORPS UNIVERSITY LIBRARIES
NATIONAL DEFENSE UNIVERSITY LIBRARY
NAVAL HEALTH RESEARCH CENTER WILKINS BIOMEDICAL LIBRARY
NAVAL POSTGRADUATE SCHOOL DUDLEY KNOX LIBRARY
NAVAL RESEARCH LABORATORY RUTH HOOKER RESEARCH LIBRARY
NAVAL WAR COLLEGE LIBRARY
NAVY PERSONNEL RESEARCH, STUDIES, AND TECHNOLOGY SPISHOCK LIBRARY (3)
PENTAGON LIBRARY
USAF ACADEMY LIBRARY
US COAST GUARD ACADEMY LIBRARY
US MERCHANT MARINE ACADEMY BLAND LIBRARY
US MILITARY ACADEMY AT WEST POINT LIBRARY
US NAVAL ACADEMY NIMITZ LIBRARY