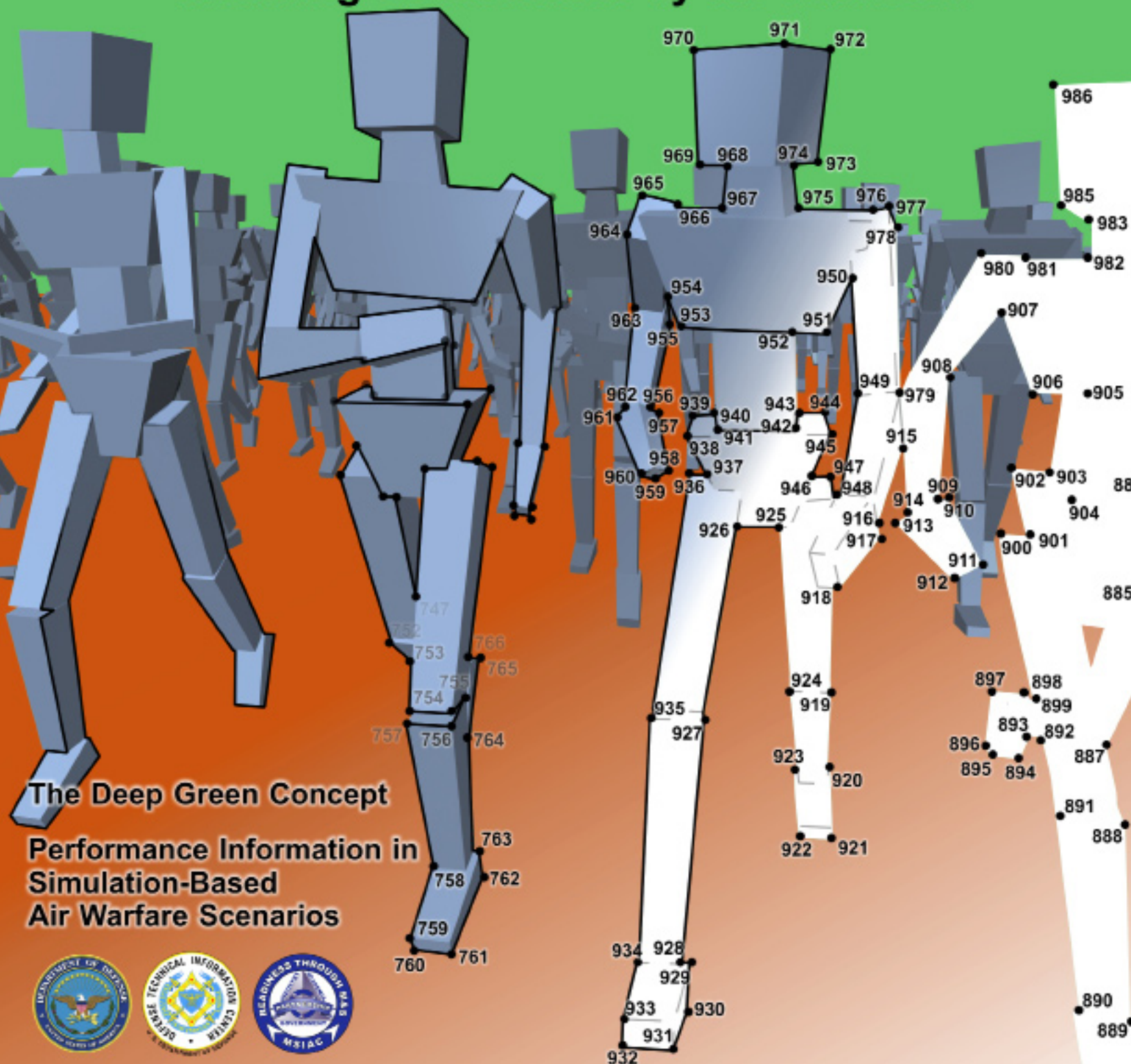
MSIAC JOURNAL

Special Edition Nov 2008 Volume 3 Issue 3

Modeling & Simulation Information Analysis Center

Connecting the Dots:

Modeling & Simulation by the Numbers



The Deep Green Concept

Performance Information in Simulation-Based Air Warfare Scenarios

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
MSIAC Journal, Volume 3, Issue 3, November 2008	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Modeling and Simulation Information Analysis Center (MSIAC),1901 N. Beauregard Street Suite 500,Alexandria,VA,22311**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**24**	

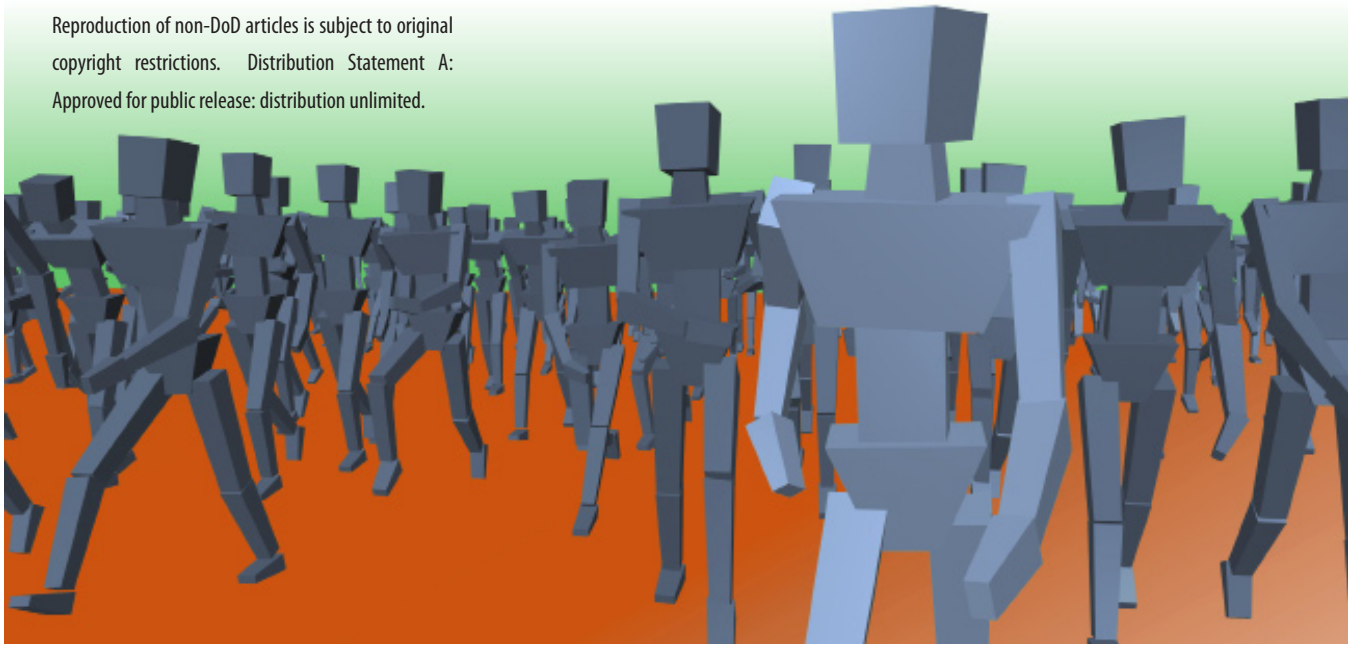
Table of Contents

The Modeling and Simulation Information Analysis Center (MSIAC) Journal is now available as an automated service. Simply send an email to journal-subscribe@lists.dod-msiac.org to be added to our mailing list. This list is for the MSIAC Journal only and will not be used for any other purpose.

To unsubscribe please send an email to journal-unsubscribe@lists.dod-msiac.org.

Please note that it is not necessary to resubscribe each month. If you would like to submit a technical paper for consideration for our Journal, please email it along with contact information to msiachelpdesk@dod-msiac.org.

The appearance of an article in the MSIAC Journal does not constitute an endorsement by the DoD, the Modeling and Simulation Information Analysis Center (MSIAC), the Modeling and Simulation Coordination Office (M&S CO), or the Defense Technical Information Center (DTIC), or any of the affiliated government contractors.



From the Director

CONNECT THE DOTS. This child's game may strike you as a strange theme for a scholarly M&S journal, but I think careful adult reflection proves otherwise. To set the stage for your enjoyment of this special issue of the MSIAC Journal for I/ITSEC 2008, please think of M&S as we do at the MSIAC: an organized approach for investigating, interpreting, understanding, and practicing real-world behaviors, situations, and processes. For success, M&S needs to represent current or anticipated systems, people, processes, and environments in a consistent fashion to develop insights, specifications, predictions, and skills. But how do we do this?

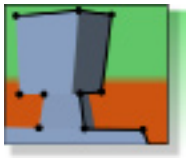
To start the process of investigating behaviors, situations, and processes, we form notions – or ideas – of how things work. These ideas are never complete, but are "abstractions" describing each part of a generally large system, system of systems, or family of systems under consideration. Our abstractions of these parts are mental "dots" that act as place holders representing our concepts of these parts. These dots are the important basics of understanding.

In the child's game, the image of the whole slowly emerges from the parts as we draw line segments between the points – that is, *connect the dots*. Similarly, in M&S, understanding of the whole system slowly emerges from the understanding of the parts – mental dots – as we describe the interactions and interfaces between these parts. The process of conceptualizing, abstracting, and modeling these connections is the next important step that builds on the basics. However, in our current state of development in the field of M&S, this step of making the connections between the dots is still more of an art than a science.

This special issue of the MSIAC Journal for I/ITSEC 2008 presents two extremely interesting articles exploring new ways to *connect the dots* by applying M&S to a range of evolving issues. The paper by COL Surdu explores what DARPA does best – looking at high risk, high payoff applications of new technology. The Deep Green concept illustrates a future where planning and operations capabilities are supported by integrating emerging M&S technologies that describe interactions in an innovative way. The article by Radtke et al. explores new paths to evaluating and enhancing the effectiveness of existing training simulations through improved operator/system interactions for after action reviews. The methods proposed and evaluated should apply as well to a wide range of simulation applications employed by experimentation, testing, and other communities enabled by M&S.

Both of these papers investigate interactions – connections – and offer ideas and approaches to improving the development and execution of M&S. Let me close by asking you to keep in mind one additional thought as you read these articles: the last dot that you might need is the MSIAC itself – contact us for personalized M&S support to *connect YOUR dots*.

Dane Mullenix, MSIAC Director



The Deep Green Concept

Written by: *COL John R. Surdu, Ph.D. Defense Advanced Research Projects Agency*
Kevin Kittka, Science and Technology Associates, Inc.

Keywords: Discrete event simulation, command and control, decision support systems, qualitative reasoning, planning

Abstract

The Deep Green concept is an innovative approach to using simulation to support ongoing military operations while they are being conducted. The basic approach is to maintain a state space graph of possible future states. Software agents use information on the trajectory of the ongoing operation, vice a priori staff estimates as to how the battle might unfold, as well as simulation technologies, to assess the likelihood of reaching some set of possible future states. The likelihood, utility, and flexibility of possible future nodes in the state space graph are computed and evaluated to focus the planning efforts. This notion is called anticipatory planning and involves the generation of options (either automated or semi-automated) ahead of "real time," before the options are needed. In addition, the Deep Green concept provides mechanisms for adaptive execution, which can be described as "late binding," or choosing a branch in the state space graph at the last moment to maintain flexibility. By using information acquired from the ongoing operation, rather than assumptions made during the planning phase, commanders and staffs can make more informed choices and focus on building options for futures that are becoming more likely. This paper will describe the Deep Green concept in detail.

> *"Key to the art of command is not to select the best course of action, but to select one that has the most, and best, options at the last minute. A good enemy is prepared for your best COA. You can't append surprise and deception to the best COA." --GEN (Retired) Richard Cavazos* [1]

1. OVERALL VISION FOR DEEP GREEN

In a military operational environment the only invariant is constant change, particularly the situation and goals. Under uncertain and time-critical conditions, it is important for commanders to have the ability to rapidly understand the unfolding trajectory of the operation and generate options quickly. More importantly, however, in modern warfare, it is important for the commander to be able to proactively generate options well in advance of when those options are needed rather than generate options reactively as the situation forces him off the plan. In this situation, it is much more important for the commander to have options than to have planned the optimum course of action in fine detail. Robust plans are those that provide not just good outcomes but maximum flexibility to adapt to unforeseen or unexpected situations.

The Defense Advanced Research Projects Agency (DARPA) has recently release a broad area announcement (BAA), 07-56 Solicitation [2] for a battle command technology program, called Deep Green. Going beyond IBM's "Deep Blue"[3] Supercomputer for Chess, Deep Green is meant to be a commander-driven technology, rather than on building technologies to remove the commander. The Deep Green program has the goal of providing tactical commanders a technology to:

- generate and analyze options quickly, including generating the many possible futures that may result from a combination of friendly, enemy, and other courses of action;
- use information from the current operation to assess which futures are becoming more likely in order to focus the development of more branches and sequels; and
- make decisions cognizant of the second- and third-order effects of those decisions.

Deep Green is composed of tools to help the commander rapidly generate courses of action (options) through multimodal sketch and speech recognition technologies. Deep Green will develop technologies to help the commander create courses of action (options), fill in details for the commander, evaluate the options, develop alternatives, and evaluate the impact of decisions on other parts of the plan. (See Figure 1.) The permutations of these option sketches for all sides and forces are assembled and passed to a new kind of combat model which generates many qualitatively different possible futures. These possible futures are organized into a graph-like structure. The commander can explore the space of possible futures, conducting "what-if" drills and generating branch and sequel options. Deep Green will take information from the ongoing, current operation to estimate the likelihood that the various possible futures may occur. Using this information, Deep Green will prune futures that are becoming very improbable and ask the commander to generate options for futures that are becoming more likely. In this way, Deep Green will ensure that the commander rarely reaches a

point in the operation at which he has no options. This will keep the enemy firmly inside our decision cycle – even an enemy that does not subscribe to a formal decision making process.

We assert that the venerable Observe-Orient-Decide-Act (OODA) loops [4] no longer viable for an information-age military. Deep Green creates a new OODA loop paradigm. When something occurs that requires the commander's attention or a decision, options are immediately available. When the planning and execution monitoring components of Deep Green mature, the planning staff will be working with semi-automated tools to generate and analyze courses of action ahead of the operation while the command concentrates on the Decide phase. By focusing on creating options ahead of the real operation rather than repairing the plan, Deep Green will allow commanders to be proactive instead of reactive in dealing with the enemy.

Deep Green was inspired by two concepts: anticipatory planning and adaptive execution. **Anticipatory planning** can be described colloquially as "you know you're going to re-plan anyway, so why not re-plan ahead of time?" This drives the notion of generating options and futures before they are needed. To some extent Deep Green will trade depth for breadth. Today commanders plan a small number of options very deeply, i.e., all the way to the end of execution in great detail. Most of these deep plans are discarded once the plan goes awry. Sometime the commander and staff are unable to recognize that the plan is broken or is becoming broken. They are often unable to divorce themselves from the plan in order to seek new affordances based on the current state of the operation. By identifying the trajectory of the operation and focusing the commander and staff where to build (perhaps less deep) plans, the commander will have a broader set of options available at any time. This leads to the concept of **adaptive execution**[5], which is similar to the AI planning concept of late binding. Adaptive execution intends to make decisions at the last moment in order to maintain flexibility to adapt to updated trajectories of the operation.

2. BASIC SYSTEM ARCHITECTURE

2.1. Commander's Associate

The Commander's Associate has two major sub¬components, Sketch to Plan and Sketch to Decide. (See Figure 2.) The two components are discussed separately because in an open, modular architecture, it is envisioned that one or the other must be able to be replaced with new technologies over time without disrupting the entire system. A goal of the Deep Green program is to de-

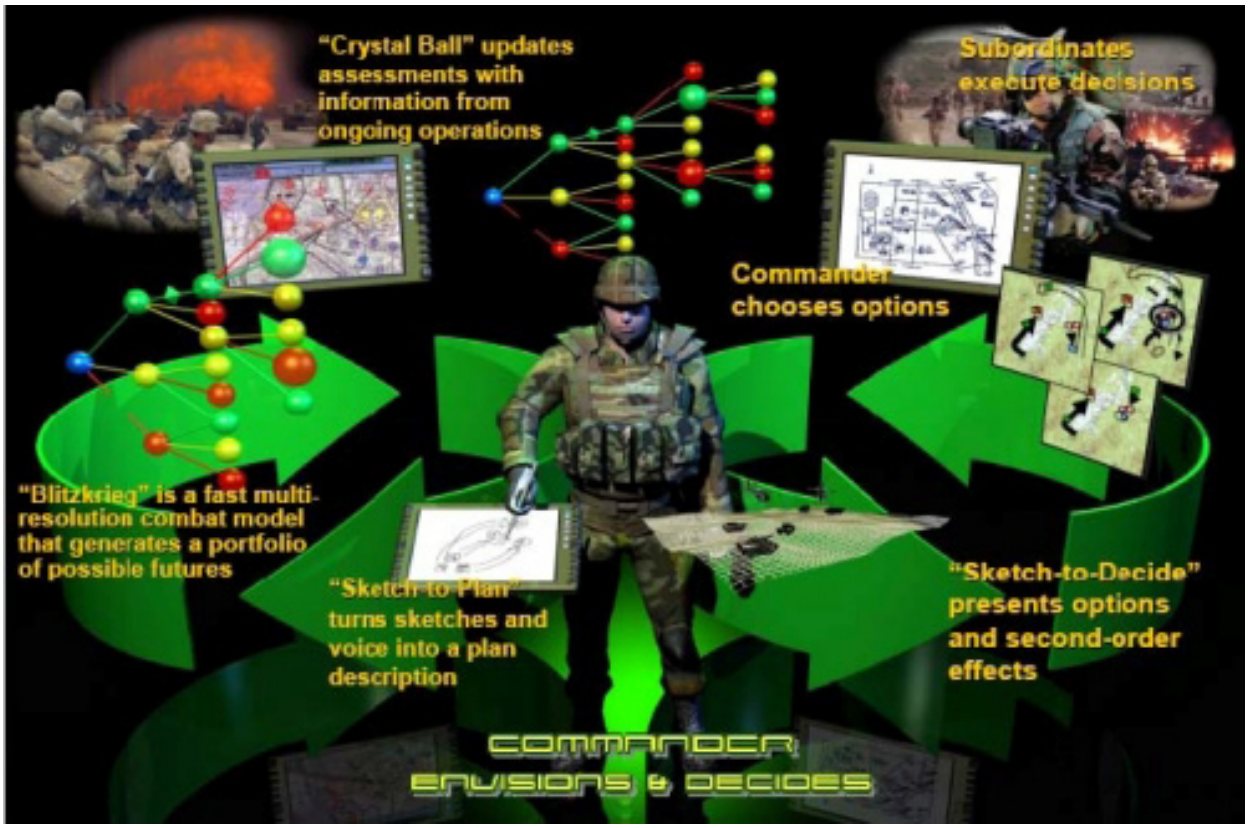


Figure 1: Operational Concept for Deep Green

velop and apply computer software technologies to develop a Commander's Associate that automatically converts the commander's hand-drawn sketch with accompanying speech of his intent into a Course of Action (COA) at the brigade level. The Commander's Associate must facilitate option generation, "what-if" drills, and rapid decision making.

Sketch to Plan

This component provides the commander the ability to generate quickly qualitative, coarse-grained COA sketches that the computer can interpret. Sketch to Plan will be multi-modal (both sketching and speech) and interactive. The computer will watch the sketch being drawn and listen for key words that indicate sequence, time, intent, etc. as the commander is creating the sketch. Sketch to Plan will induce both a plan and the commander's intent from the sketch and speech. Unlike other approaches that are optimized around machine interpretations [6] (i.e. constraining the sketching method to drag-and-drop modalities, forcing the human to learn the computer's 'language' to some extent), Sketch to Plan is optimized around the user free-hand sketching options over a map. In addition, the Sketch to Plan component must be imbued with enough domain knowledge that it knows what it doesn't know and can ask the user a small set of clarifying questions until it understands the sketch and can use it to initialize a combat model.

The sketch Recognizer converts a free-hand set of strokes, combined with speech, into a set of military objects, such as units and graphical control measures (MIL STD 2525b [7] and STANAG 2019 APP-6A [8]). The plan inducer has the challenge of inducing the commander's plan and intent for the recognized "bag of symbols." We envision a detail adding planner within Sketch of Plan that adds details to the commander-generated option so that it can be modeled by Blitzkrieg. Finally, the dialog generator helps Sketch to Plan understand the commander's option by formulating clarifying questions when necessary.

Sketch To Decide

When the commander is asked for a decision, Sketch to Decide will allow him/her to explore the future space to gain an appreciation for the ramifications of a choice. It is envisioned as similar to a comic strip with branch points that correspond to branch points in the futures graph. Scott McCloud [9] asserts that the idea of a comic in which the readers get to make a choice at the branch points is today "exotic" but may well become common in the future. Since the 1970s (and perhaps earlier), there have been novels and game books in which the reader is asked to make a decision and then is directed to a different page or paragraph, depending on the choice made, such as the 1980's children's Choose Your Own Adventure gamebook series or the DVD movie Clue based on the board game Clue as examples . Recently Forbus has explored the idea of a comic graph [10]. The idea here is the same: the user gets to choose which path to follow at a branch point. One can imagine the commander exploring the future space to understand how his courses of action may play out and identifying the critical branch (decision) points.

Sketch to Decide is designed to allow the user to "see the future," but this capability must be developed with care to prevent confusing the decision space. Humans are notoriously bad at thinking through probabilistic choices and even more so when there are competing outcome utilities. At each branch point, there are multiple decision dimensions/utilities that have to be considered, such as likelihood, risk, utility, resource usage, etc. In addition, the abstract nature of the state and the uncertainty of predictions, locations of units, etc. must be portrayed intuitively. Therefore, at any "frame" in the Sketch to Decide graph, the user can perform Sketch to Plan actions, allowing the commander to conduct "what-if" drills wherever he wants in the future space. The user is going to need a lot of help in evaluating these options, especially because they are probabilistically weighted. By presenting decisions early and allowing the commander to explore the future space, Sketch to Decide supports adaptive execution, allowing the commander to make decisions when they are needed, rather than committing too early.

2.2. Blitzkrieg

Blitzkrieg is the simulation component of Deep Green. It is used to generate the possible futures that result from a set of plans (one plan for each side/force in the operation). Besides being very fast (the blitz in Blitzkrieg), it is designed to generate a broad set of possible futures. These futures should be feasible, even if not expected by human users. Over time, Blitzkrieg should learn to be a better predictor of possible futures, based on presented options. Blitzkrieg identifies branch points, predicts the range of possible outcomes, predicts the likelihood of each outcome, and then continues to simulate along each path/trajectory. Gilmer and Sullivan provide an example of a possible implementation of this idea [11] in which they determine branch points and continue to simulate along multiple paths. Blitzkrieg should reflect out-of-the-box thinking, rather than merely generating hundreds or thousands of "Monte Carlo" runs of a stochastic model and binning the outputs [12]. This will require an innovative hybrid of qualitative and quantitative technologies.

As an example, two forces may collide with each other. The collision may be predicted with some sort of analytical

model that accounts for non-determinism in rate of march of the forces. Qualitatively there are a number of possible outcomes of this collision: one side or the other may get quickly defeated, one side may begin to lose and withdraw, the two forces might avoid each other and continue on their way, both sides may choose not to engage each other, or both sides may become involved in an attrition slug-fest, etc. Quantitative models, such as Lanchester equations [13] or the Qualitative Judgment Model [14] might then be used to determine the relative likelihood of these various outcomes. Perhaps heuristic methods might be used instead of or in addition to these quantitative models. For instance, a fuzzy rule base might be used that takes into account aggressiveness of the opponents, their relative

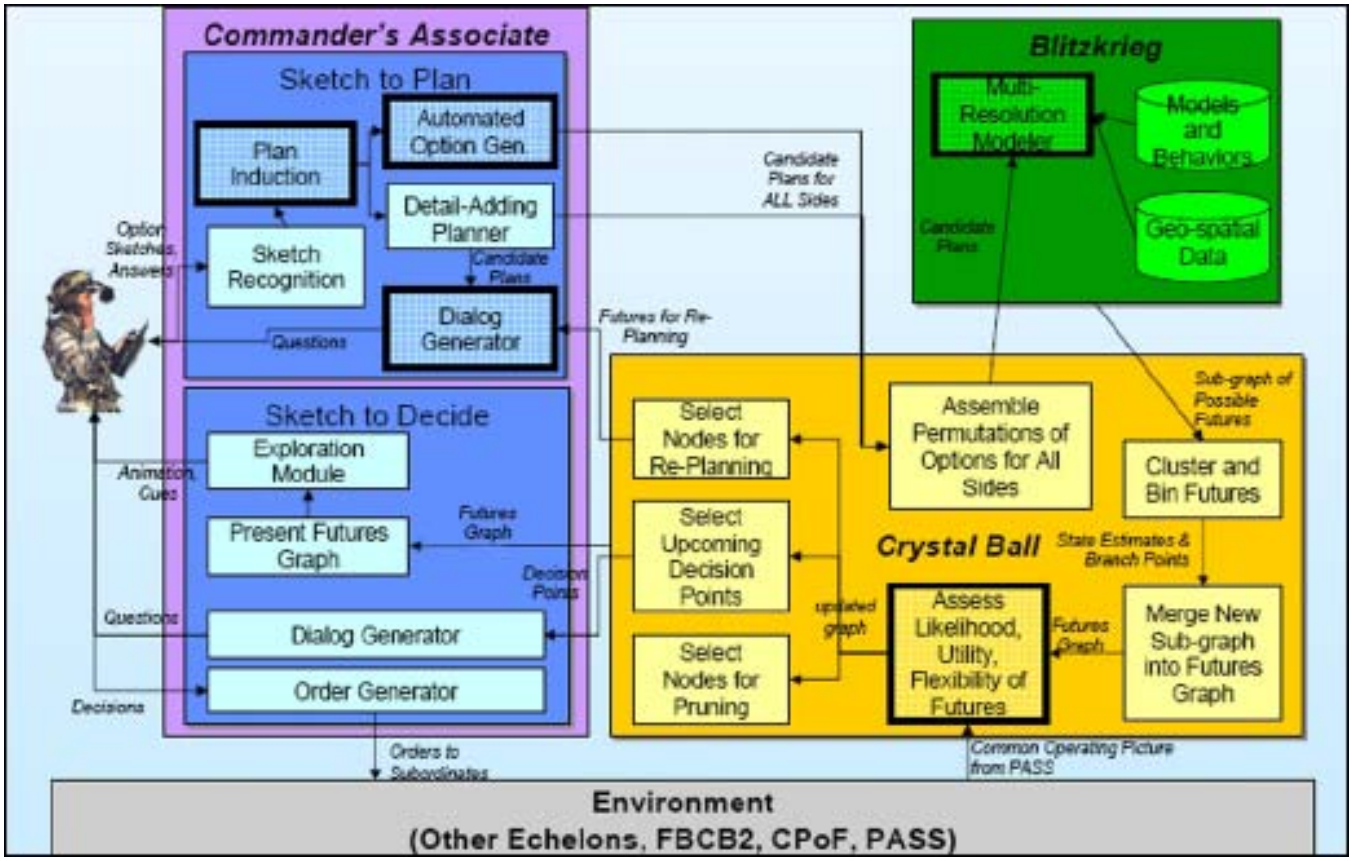


Figure 2: Architectural Overview of Deep Green

strengths, etc.

In warfare, all the players can be potentially moving at the same time, so predicting when these forces will meet, separate, etc. is challenging. The conditions of these meetings may, in fact, also impact the prediction of outcomes described in the previous paragraph. Continuing with this scenario, due to the non-deterministic nature of each side's movement, speeds could indicate some likelihood that one side or the other would reach a key piece of terrain first. In this case, the force that arrived first might have an advantage in the ensuing engagement. If, on the other hand, the force that arrives first is in an exposed position, such as being in the middle of a river crossing or out in the open, the other side might have an advantage.

The current war has many non-kinetic aspects and involves paramilitary forces, terrorists, and masses of civilians on the "battlefield." Blitzkrieg, and in fact all of Deep Green, must support the full spectrum of military operations, from mid-intensity combat to operations other than war, perhaps all occurring simultaneously in a three-block war context [15]. We believe that the combination of these qualitative and quantitative methods will allow Blitzkrieg and Deep Green to better support full spectrum operations. The impacts of medics and food distribution in local villages, the destruction of culturally significant sites, morale, leadership, and cohesion perhaps are best represented qualitatively, rather than quantitatively.

Today's class of combat models requires detailed terrain databases in order to function properly. Blitzkrieg will use more qualitative terrain representations. Commanders do not reason on the stem spacing and diameter of trees at breast height, vertical cone index of soil, or whether a

particular area is composed of sandy clay loam. They reason about maneuver corridors, key terrain, and points of dominance. Of course, we do not want to "dumb down" Blitzkrieg to the extent that it provides little additional rigor than would an average human, but the right balance needs to be struck. At the same time, the creation of the abstract, qualitative terrain representation should be based on the same detailed terrain representation used in our current class of simulations, such as the OneSAF Objective System [16] Objective Terrain Format [17], and generate the more abstract terrain needed by Blitzkrieg in an automated fashion.

2.3. Crystal Ball

Crystal Ball serves several functions. First, it controls the operation of Blitzkrieg in generating futures. Second, it takes information from the ongoing operation and updates the likelihood metrics associated with possible futures. Third, it uses those updated likelihood metrics to prune parts of the futures graph and nominate futures at which the commander should generate additional options and invokes Sketch to Plan. Finally, it identifies upcoming decision points and invokes Sketch to Decide. While Crystal Ball has a moderate role prior to execution, it is the backbone of the system during execution.

Prior to Execution

During pre-operations planning, Crystal Ball receives options from Sketch to Plan for all sides and forces. These options are generated by humans. Crystal ball assembles the permutations of plans and sends them to Blitzkrieg to generate the possible futures that result from each permutation. If the commander uses Sketch to Decide to inject branches and sequels into this process, Blitzkrieg will make additional runs. Blitzkrieg returns sub-graphs of possible futures and branch points to Crystal Ball with annotations as to Blitzkrieg's a priori estimate of the likelihood of these options. Another function of Crystal Ball is to merge these sub-graphs so the futures that are qualitatively the same (regardless of which permutation of options generated them) are combined. This reduces the complexity of the future space, helps refine the list of critical branch points in the future space, and makes Crystal Ball's during-execution job easier.

Crystal Ball also generates two additional metrics associated with the possible futures: value/utility and flexibility. Utility is a rating of how good the future is with respect to the goal of the operation. Utility cannot be based completely on some a priori estimate of "board position," casualty rates, etc. "Board positions" are really a measure of the location of entities with respect to key terrain, the objective, etc., but what constitutes key terrain can often be a function of the mission. Flexibility is a measure of how many branches from a future lead toward better utility. Most commanders would rather have choices than only one good path. If the battle is moving toward nodes with little flexibility, this indicates that the plan is "brittle" and perhaps can be easily derailed by enemy action – or our own mis-actions.

During Execution

Once the operation is underway, Crystal Ball will get information about the ongoing operation from the battle command systems, such as FBCB2 [18], CPoF [19], or the publish and subscribe services (PASS) [20] of ABCS 6.4+. For forces other than BLUE, this information is largely location and perhaps strength information fused from various intelligence sources. (This information fusion is not a part of Deep Green's objectives; Deep Green assumes the information it gets is the best available.) For BLUE forces this information will include information about location and strength, but also potentially information about logistics status, velocity, etc. Crystal Ball uses this information about the current operation to update the likelihood estimates of the many possible futures. Having done that, Crystal Ball can compare the likelihood, utility, and flexibility and estimate which futures are likely to occur that have little value or flexibility. Crystal Ball will use this estimate to nominate to the commander futures at which he/she should focus some planning effort to build additional options/branches. If the commander reaches a future for which no options have been developed, he/she has been surprised and the enemy is now operating inside his/her decision cycle. Crystal Ball will identify the trajectory of the operation in time to allow the commander to generate options before they are needed. Crystal Ball will also use this information and additional heuristics to nominate futures for pruning from the graph and to identify decision points to send to Sketch to Decide. Pruning, however, will not be based purely on likelihood, but also on attributes such as risk to the operation.

2.4. Automated Option Generation

The focus of Deep Green is on tools to help the commander (and staff) generate options quickly. Leaders from the field generally do not want machine-generated courses of action. Nevertheless, under Deep Green, we intend to sponsor a small set of modest efforts to generate options automatically. The long-term vision of Deep Green is for options to be generated by both the commander and the computer. Initially we expect the machine generation of options to be centered on making clever "mutations" of the human-generated options to increase the breadth of the futures generated. This highlights the need for Sketch to

Plan to induce the commander's intent from the free-hand sketches. Any options generated by the computer should feasibly meet the commander's intent.

2.5. State Space Graph

Throughout this discussion of Deep Green we have mentioned the "state space graph." We are still very early in the development of Deep Green; in fact, by the time this paper is published we will have just selected performers for Deep Green. We envision the collaboration of Blitzkrieg and Crystal Ball creating and maintaining a graph of possible futures. The tasks assigned to Crystal Ball sound like a hybrid of Markov technologies, such as hidden Markov models, Markov Chain Monte Carlo, Markov and Partially Observable Markov Decision Processes, and Bayesian technologies, such as continuous bayes networks, Gaussian, Inference, and Clustering/Join Trees [21]. We, therefore, envision the data structure of the state space graph also being a hybrid of these representations. One can envision Blitzkrieg adding nodes to the graph and Crystal Ball updating, and in some cases pruning, the graph. Conceptually, this would appear like the movement of an amoeba, where the human-generated options cause Blitzkrieg to shoot out pseudopodia. In preparation for initiating Deep Green, we commissioned a study to look at existing planning languages in the AI community and the military and identify the necessary and sufficient data elements for this state space graph. That report will be published in the future.

3. FUNDAMENTAL SHIFT AWAY FROM THE TRADITIONAL OODA PARADIGM

Figure 3: Multiple OO's, One DA Loop Processes Sketch to Decide

The OODA loop concept [22] was first introduced by Col John Boyd, U.S. Air Force fighter pilot ace, in 1986 in his presentation entitled "Patterns of Conflict" (POC). (See Figure 3) Since then there have been many variations of this process. The venerable Observe-Orient-Decide-Act (OODA) loop is no longer viable for an information-age military. Previous work has centered on speeding up the overall loop or developing technologies that work within a single phase of that loop. Today, when the plan goes awry, we go into a reactive mode, in which we create courses of action, analyze them, and then choose.

Deep Green creates a new OODA loop paradigm. (See Figure 4) Observe (execution monitoring) and Orient (options generation and analysis) phases run continuously and are constantly building options based on the current operation and making predictions as to the direction the operation is taking. When something occurs that requires the commander's attention or a decision, proactive options are immediately available. Ideally, the OO part of OODA is

Figure 4: The OODA Loop

done many times prior to the time when the commander must decide. When the planning and execution monitoring components of Deep Green mature, the planning staff will be working with semi-automated tools to generate and analyze courses of action ahead of the operation while the command concentrates on the Decide phase. By focusing on creating options ahead of the real operation rather than repairing the plan, Deep Green will allow commanders to be proactive instead of reactive in dealing with the enemy.

4. DEEP GREEN IN OPERATION

The authors have described the high-level, technical underpinnings of the Deep Green concept. It may not, however, be clear how Deep Green would function in an

operational context. The authors will resort to a bit of fiction to help convey this vision.

Imagine that a brigade headquarters is tasked to simultaneously conduct stability patrolling in an area, create and run a food distribution point, and also conduct a raid to seize a known enemy leader in the area of operation. The commander quickly sketches an option using Sketch to Plan. He then directs the intelligence officer to create some options for the enemy (using Sketch to Plan) and the operations officer to generate two more options (also using Sketch to Plan).

As the intelligence officer completes the first option for the enemy (and perhaps what he believes the indigenous population might do), Crystal Ball passes that option along with the commander's option to Blitzkrieg to generate futures. As more options are generated for all sides, Blitzkrieg generates more futures.

Later, the food distribution point has been established near a market and the combat patrol is zooming toward the suspected location of the enemy leader. Talking to a local business leader, one of the dismounted patrols gathers human intelligence that two competing warlords are planning to attack the food distribution point to seize the food and distribute it to their own followers. This is corroborated by a report from an unmanned aerial vehicle of the movement of suspected warlord vehicles departing a neighboring village toward the village with the food distribution point, due to arrive in forty minutes.

Knowing that the commander will want to know if forces of the rival warlord are also on the move, Sketch to Decide asks the Automated Option Generator to create a plan to re-task an unmanned aerial vehicle over the area where his forces are known to operate. This is presented to the intelligence officer, who approves the option.

As a result, the likelihood of the future in the futures graph in which the food distribution is attacked goes up. Worried that this attack may take place at the same time as his combat patrol is raiding the enemy leader's location, the commander sketches three options: one in which two of the stability patrols are moved to a position to support the food distribution point, with the goal of preventing the warlords from attacking; another in which one of the stability patrols assumes the raid mission and the mounted, combat patrol moves to support the food distribution point; and another in which the food distribution point closes up and returns to base. It takes less than ten minutes to sketch these options. The Detail Adding Planner fills in additional details and passes them to Blitzkrieg, which generates a number of new futures. One of these new futures indicates that the movement of the dismounted patrols spooks the suspected enemy leader who is the target of the combat patrol, and he flees. The operations officer sketches options for how they will respond if the enemy leader begins to move…

The people in the village are dependent on the food for survival. The enemy is spreading propaganda that the U.S. forces aren't committed to feeding them and that only they can help the people. Folding up the food distribution point, even for a day, will play into the hands of the enemy. Deep Green predicts a drop in friendliness of the local population if they take that option. This will impact the success of future operations and the overall mission of the U.S. Army.

Trucks that are suspected of carrying the forces of the rival warlords continue to move toward the food distribution point, so the likelihood of an attack on the food distribution point in an hour does not drop as expected. The operations officer generates options in which smart munitions are used to stop these vehicles. The movement of the dismounted patrols is slower than expected, because of heavy traffic on market day. The likelihood of them intercepting the warlord's forces or getting between them and the food distribution point goes down.

While all this planning is occurring, Sketch to Decide recognizes that the time has arrived for the commander to make a decision whether to send the mounted patrol to the food distribution point or stay the course with the dismounted patrols or the mounted patrol will not be able to reach the food distribution point in time. This decision point is presented to the commander in time to let him explore the future space and get a feel for second- and third-order effects and unintended consequences.

In the meantime, the intelligence officer has picked up reports of a possible attack on one of the brigade's platoon patrol bases within the city in the next week, and the operations officer begins to sketch options to head off the attack, to respond if attacked, etc.

Just as one of the rival warlords nears the food distribution point and is confronted by one of the dismounted patrols, the enemy leader flees the building that was the target of the mounted combat patrol…

5. SUMMARY

We are just getting started! The selection of performers for Deep Green was completed in February 2008. We anticipate that they will be on contract in late April and begin work. The first major milestone will be twelve months later. Deep Green will provide technology to break the OODA paradigm and will enable the rapid construction of sophisticated planning and execution systems using existing technologies. The overall objective will be an open and scalable battle command decision support architecture

that interleaves anticipatory planning and adaptive execution to stay inside the enemy's decision cycle. Deep Green will provide an implementation framework to enable rapid technology insertion into battle command systems today and in the future. When successful, we will build a revolutionary decision support system that will allow us to defeat peer competitors.

6. REFERENCES

[1] Cavazos, GEN (Retired) Richard, USA. (1993). Title unknown. Source from transcript of speech given at US Army Command and General Staff College (CGSC) by GEN Cavazos. (Thanks to Ryan Paterson and his Team @ Praevius Group for bringing this quote to our attention.)

[2] Surdu, John R., COL. (2007). Deep Green Broad Agency Announcement No. 07-56. Defense Advanced Research Projects Agency (DARPA), Information Processing Technology Office (IPTO), Website: http://www.darpa.mil/ipto/solicitations

[3] Unknown. (1997). Kasparov Vs. DEEPBLUE: The Rematch – Overview. IBM. Website http://www.research.ibm.com/deepblue/

[4] John R. Boyd, C. (1986). "Patterns of Conflict." 196. http://www.d-n-i.net/boyd/pdf/poc.pdf

[5] Brigadier General Huba Wass de Czege, U. A., Retired and U. A. Major Jacob Biever (1998). "Optimizing Future Battle Command Technologies." Military Review(March/April 1998): 6.

[6] Forbus, K. D., J. Usher, et al. (2003). Sketching for Military Courses of Action Diagrams. IUI'03. Miami, Florida, USA, ACM.

[7] Unknown (1999). "Department of Defense Interface Standard: Common Warfighting Symbology - MIL¬STD-2525B ": 556. http://assist.daps.dla.mil/docimages/0001/52/27/2525B. PD5

[8] Unknown (2000). "NATO Standardization Agreement 2019: APP6A Military Symbols for Land Based Systems." http://www.mapsymbs.com/app-6ahandbook.zip

[9] McCloud, S. (1994). Understanding Comics: The Invisible Art. New York, NY, HarperCollins Publishers, Inc.

[10] Forbus, K. D., J. Usher, et al. (2003). Qualitative Spatial Reasoning about Sketch Maps. Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence, Acapulco, Mexico, American Association for Artificial Intelligence Press.

[11] Gilmer, J. B., Jr. and F. J. Sullivan (2005). "Issues in Event Analysis for Recursive Simulation." Proceedings of the 2005 Winter Simulation Conference: 8.

[12] Law, A. M. and W. D. Kelton (2000). Simulation modeling and analysis. United States of America, The McGraw-Hill Companies, Inc.

[13] Unknown "Lanchester Equations and Scoring Systems." Website http://www.rand.org/pubs/monograph_reports/MR638/a pp.html

[14] Dupuy, T. N. (1995). "Attrition: Forecasting Battle Casualties and Equipment Losses in Modern War." Nova Publications. Falls Church, VA.

[15] Krulak, GEN. C. C. (1999). The Strategic Corporal: Leadership in the Three Block War. Marines Magazine. Website http://www.au.af.mil/au/awc/awcgate/usmc/strategic_c orporal.htm

[16] Unknown. (2007). "OneSAF Public Site." Retrieved 9-11-07, 2007, from http://www.onesaf.net/community/.

[17] Unknown. (2007). "Army Digital Terrain Catalog." Retrieved 9-11-07, 2007, from http://www.tec.army.mil/fact_sheet/ADTC.pdf.

[18] Unknown. (1998, September 12, 1998 6:35:55 AM). "Force XXI Battle Command, Brigade-and-Below (FBCB2)." 2007, from http://www.fas.org/man/dod¬101/sys/land/fbcb2.htm.

[19] Unknown. (2007). "PM Battle Command: Command Post Of the Future (CPOF)." 2007, from http://peoc3t.monmouth.army.mil/battlecommand/bc_C POF.html.

[20] Unknown. (2007). "PM Battle Command: Maneuver Control System (MCS)." 2007, from http://peoc3t.monmouth.army.mil/battlecommand/bc_m cs_2.html.

[21] Russell, S. and P. Norvig (2003). Artificial Intelligence: A Modern Approach. (2nd Ed) Upper Saddle River, NJ, Pearson Education, Inc.

[22] Boyd, Col. J. R. (1986). "Patterns of Conflict." (Pg 132) 196 pages from http://www.d-n¬i.net/boyd/pdf/poc.pdf

Biography

COL John R. Surdu, Ph.D., U.S. ARMY, Program Manager, DARPA Information Processing Technology Office. (*john.surdu@us.army.mil*) COL "Buck" Surdu was commissioned a second lieutenant of infantry after graduating from the United States Military Academy in (1985).

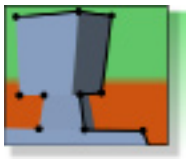
COL Surdu has worked as a research scientist and team leader at the Army Research Laboratory, focusing on unique uses for virtual reality technologies for command and control applications. As a senior research scientist at the Information Technology and Operations Center, he directed several applied research efforts. From 2003-2006 COL Surdu has been the Product Manager for the One SAF program office.

In addition to a Bachelor of Science degree in computer science from the United States Military Academy, in (1985) COL Surdu earned a Master of Business Administration degree from Columbus State University. In (1990), COL Surdu earned Master of Science degree in computer science from Florida State University (focusing on artificial intelligence in 1995). He finalized his formal education with a doctoral degree in computer science from Texas A&M University in (2000) (focusing on simulation technology and its applications to command and control).

Kevin Kittka, Program Manager, Science and Technology Associates, Inc., (*kkittka@stassociates.com*) has a B.S. from The Pennsylvania State University and a M.S. from The University of Maryland University College. Mr. Kittka has provided consulting support to NAVSEA for 8 years and Scientific, Engineering, Technical, and Administrative Support to the Defense Advanced Research Projects Agency for the past 19 years. His research focus is on Information Systems. Currently a member of AAAI and an associate member of IEEE.

Note: This document is Approved for Public Release, Distribution Unlimited (Case # 10307).

This paper originally appeared in the Proceedings of the Military Modelling and Simulation Symposium 2008, part of SpringSim Multiconference held in Ottawa, April 2008. It has been reproduced with permission from The International Society for Modelling and Simulation.



Integrating and Presenting Performance Information in Simulation-Based Air Warfare Scenarios

Written by: *Paul H. Radtke & Joan H. Johnston, Naval Air Warfare Center Training Systems Division Orlando, FL*
Elizabeth Biddle, The Boeing Company, Training Systems & Services, Orlando, FL
Thomas F. Carolan, Alion Science & Technology MA&D Operation, Boulder, CO

ABSTRACT

Simulation-based tactical training exercises are ideal settings in which to evaluate performance. The capability to record the second-by-second behavior of participants, the state of supporting equipment, and the location of entities in the problem provides an opportunity to verify team and individual proficiency, and to identify root cause of substandard performance. However, responsibility for determining cause and effect in tactical scenarios is typically left to the expert instructor. In dynamic, fast-paced warfare areas, such as air-to-air combat, the burden on the unaided expert instructor to monitor, record, and assess the interactions and circumstances that determine mission success, is substantial. This is an area where appropriate technology might help the instructor to improve the evaluation of performance.

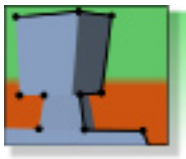
The Debriefing Distributed Simulation Based Exercises project (DDSBE), an ONR-sponsored 6.3 research and development project, tested alternative technologies for collecting and integrating performance information to aid in the preparation and delivery of post-scenario after action reviews (AARs). The project's objective was to provide the information that instructors need, when needed, in a form that supports rapid evaluation. This paper presents a comparison of different performance data collection, analysis, and debriefing systems, and the performance information they make available to instructors in the context of two distributed training research systems. The first system, built to support the DDSBE research effort, analyzed the performance of two E-2C Naval Flight Officers (NFOs) and F/A-18 Sweep Lead during an air-to-air engagement. Human observers and an automated data collection component collected performance data. The second system, a two-ship F/A-18 simulation built to support training research by The Boeing Company, collected and analyzed performance data for tasks performed by the Escort Lead and Strike Lead during an engagement. The paper presents and compares methods for integrating and presenting the multiple streams of performance information available to the instructor.

INTRODUCTION

Recent advances in modeling and simulation (M&S) have greatly expanded the opportunity to conduct multi-platform distributed simulation-based training exercises. For example, advances in M&S interoperability permit Navy Fleet Synthetic Training-Joint (FST-J) exercises to be conducted more quickly, and at significantly lower cost. In March 2006, US Navy, Air Force, Army, and coalition partners participated in a 72 hour FST-J exercise that would have taken over two weeks to conduct just three years ago (Glassburn, 2006). The Navy plans to increase the frequency of such exercises (Jean, 2006). However, this increased demand also results in an increased demand for evaluators who can deliver accurate estimates of mission readiness. Currently, this is a labor intensive (and costly) activity because simulators typically lack embedded tools for automated human performance assessment, diagnosis, and debrief/after action review (AAR). In dynamic, fast-paced warfare areas, such as air-to-air combat, the burden on the expert instructor is substantial. The instructor must monitor, record, and assess the actions and interactions of a large group of performers working on a rapidly changing problem, in which even small mistakes can determine mission success or failure. These tasks are made more complex and time consuming during distributed mission training exercises, in which many teams across different platforms train together but with no face-to-face interactions between instructors and training teams (Neville, Fowlkes, Milham, Merket, Bergondy, Walwanis, & Strini, 2001).

Improving the embedded assessment capabilities of distributed simulation-based training was the focus of an Office of Naval Research (ONR) sponsored program titled "Debriefing Distributed Simulation-Based Exercises" (DDSBE; Johnston, Radtke, Van Duyne, Stretton, Freeman, & Bilazarian, 2004). The DDSBE program developed M&S technologies that can mitigate the added workload of obtaining mission readiness assessments based on objective assessments of combat team and multi-team performance. Technologies were developed that record the moment-by-moment actions of team members, the state of supporting equipment, the location of entities in the problem to verify team and individual proficiency, and the root causes of substandard performance. The DDSBE program tested alternative technologies for collecting and integrating team performance information to aid in the preparation and delivery of post-scenario AARs. The project's objective was to provide the information that instructors need, when needed, in a form that supports rapid evaluation.

The purpose of this paper is to present and compare methods for integrating and presenting multiple streams of performance information available to the instructor. In this paper we compare strategies for performance data collection, analysis, and debriefing systems, and the per-



Integrating and Presenting Performance Information in Simulation-Based Air Warfare Scenarios

formance information they make available to instructors in the context of two different distributed training systems. The first system, built to support the DDSBE research effort, analyzed the performance of the E-2C Naval Flight Officers (NFOs) and the F/A-18 Sweep Lead during an air-to-air engagement. Performance data was collected by human observers and an automated data collection component. The second system was built to augment the DDSBE research with a focus on the data collection, analysis, and presentation of tasks performed by the F/A-18 team, comprised of the Escort Lead and Strike Lead, during the air-to-air engagement.

BACKGROUND

The data collected in the two experiments focused on human performance during a simulated air-to-air fighter engagement in a naval strike mission. The two data collection efforts focused on different aspects of the air-to-air engagement, but each followed the same event sequence and tactical context.

An air-to-air engagement consists of a series of voice communications, equipment manipulations, and decisions, performed by individuals or the team, and arrayed along a timeline. Satisfactory performance means performing certain procedures at the correct time, geometry, and range; using the equipment and systems effectively; making required decisions; and providing necessary information to the right person, accurately, in the prescribed format, when appropriate. The following is a description of the phases and tasks in a generic air-to-air engagement that were used to construct the scenarios, the scripted performance of the trainees used in the studies, and the associated performance measures.

For the purpose of this research, the air-to-air engagement was divided into distinct phases. The pre-commit phase began with the detection of a new, previously unidentified, aircraft by the E-2C command and control aircraft team. Based on the characteristics of the new contact – referred to as a "track" – the E-2C team was expected to assign an appropriate identification designation in the tactical data link and issues a voice report of the contact to the strike package and higher authorities. The fighter element was not expected to take any action regarding the new track except to acknowledge the communication. The fighters relied on the E-2C team to alert them when the contact becomes tactically significant. The pre-commit phase ended when the track's characteristics caused it to be designated as "hostile" and to require a response. The new designation was to be entered in the tactical data link and declared in a voice communication to the strike package.

The "hostile declaration" began the intercept phase. The E-2C vectored the escort to intercept the track. When the fighters acquired radar contact, the E-2C verified that the fighters' contact was the "bandit" in question, based on its reported altitude, bearing, and range from the fighters. The E-2C was then expected to recommend that the fighters "commit" to engage the hostile track. This began the commit phase.

During the commit phase, the E-2C monitored the engagement and passed new information to the fighters, such as any hostile aircraft maneuvers. Otherwise, the E-2C was expected to be silent and not divert the attention of the fighters as they focused on the coming engagement.

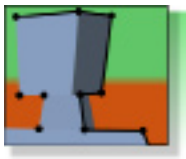
During the weapons engagement phase, the fighters attempted to hold the hostile tracks on their radar, while sorting out and targeting the tracks. They also determined the range at which they should release their weapons to minimize their vulnerability to the hostile aircrafts' weapons. The fighter pilots were expected to announce the launches with a voice communication to the E-2C.

The launch of weapons started the merge phase, during which the fighters continued to close the distance to the hostile aircraft, guided the flight of their missile, and watched for an indication that the hostile aircraft had launched a missile against them. The fighters were expected to maneuver to minimize the rate of closure while maintaining radar contact on their target until their missile could automatically track and intercept the hostile aircraft. The pilots were expected to announce this with a voice call to the E-2C. Unless the fighters were obliged to take evasive action to defeat a weapon launched at them from the hostile aircraft, the fighters continued to merge until they observed the destruction of the hostile aircraft, or confirmed that it had survived the engagement. During this phase, the E-2C operator was expected to monitor the engagement and the merge and only communicate with the fighters if there was an immediate threat.

During the post-merge phase, the fighters reported the outcome of the engagement. The E-2C provided an updated picture call to the fighters as they regrouped, prepared to reengage, or returned to their planned route. The E-2C then passed on an engagement report to the rest of the strike package and the Air Warfare commander.

DDSBE SYSTEM

The DDSBE data collection, analysis, and debrief system was developed to support an experiment focused on E-2C - F/A-18 teamwork and taskwork. This system was integrated with a simulation test bed consisting of three positions within a naval strike mission "package". Two of the positions were located on an E-2C command and control aircraft, the Air Control Officer (ACO), and the Combat Information Control Officer (CICO). These two NFOs provide



information and coordination to the other members of the mission. The third position was the Lead pilot of the F/A-18 fighter escort, or "Sweep" element, which protects the strike mission from air threats. Because the intent was to test the validity and reliability of the DDSBE system, data collection focused on the pre-scripted individual and team-level performance of the three positions. Team-level performance included the within-platform performance of the ACO and CICO, and the cross-platform teamwork of the ACO and the F/A-18 Sweep Lead.

Four scenario runs were conducted, each containing two air-to-air engagements. The first engagement involved two hostile aircraft, and the second involved a single hostile aircraft, encountering the Sweep Lead and Wingman. During two of the four scenario runs the ACO, CICO, and Sweep Lead followed pre-scripted behaviors to perform at a "nearly perfect" level. During the remaining two runs the trainees performed at a scripted "less-than-satisfactory" level. The DDSBE performance measurement plan implemented the Event-Based Approach to Training (EBAT;

Phase	Performance Measures	Automated	Manual
Pre-Commit	ACO "hooks" the new unknown track	√	
	ACO changes track ID to "Unknown Assumed Friendly"	√	
	ACO makes internal "New Track" voice report to CICO		√
	CICO "hooks" the new track	√	
	CICO changes track ID to "Unknown"	√	
	CICO enters the new track information into the tactical data link	√	
	CICO makes external "New Track" voice report to AW		√
	ACO makes external "Picture" call to Strike Package, including Sweep		√
	CICO makes internal "Aircraft Activity" voice report to ACO		√
	CICO "hooks" the track	√	
	CICO changes the track ID to "Hostile"	√	
	CICO enters the new track information into the tactical data link	√	
	ACO makes external "Picture" call to Strike Package, including Sweep		√
	SWL confirms contact report		√
	ACO recommends "Commit"		√
Commit	SWL reports "Commit"		√
	ACO "hooks" hostile track (primary hook)	√	
	ACO "hooks" Sweep lead track (secondary hook)	√	
	ACO makes internal voice report of Sweep "Commit" to CICO		√
	CICO makes external "Commit" report to AW		√
Weapon Engagement	SWL launches weapon via stick	√	
	SWL makes external "Shot" call		√
Merge	SWL makes external "Bulldog" call		√
	SWL makes external "Kill" call		√
	ACO makes internal "Kill" report to CICO		√
Post Merge	CICO acknowledges ACO's report		√
	CICO makes external "Kill" report to AW		√
	ACO makes external "Picture" call to Strike Package, including Sweep		√

Table 1. DDSBE Automated and Manual Performance Measures Collected During Air-to-Air Engagements, by Engagement Phase and Event.

Fowlkes, Dwyer, Oser, & Salas, 1998), which focuses measurement on specific, pre-identified, critical events. When these events are triggered, the participants are expected to perform particular tasks that, in turn, require that they demonstrate targeted skills, knowledge, or other types of competence. This focused approach is based on a sampling of performance and excludes analysis of events not designated to be critical.

Performance measurement relied on both automated data collection and manual input by an instructor. The Virtual Communications Assessment Tool (VCAT), a hand-held device, was used by instructors to record their observations. Two instructors observed the trainees' performance – one assigned to record the ACO and CICO, and the other assigned to observe the F/A-18 Sweep Lead. The hand-held VCAT device warned the instructor when a key or critical event was about to occur and prompted the instructor to record specific observations during the event. The information collected by the human and automated systems filled measurement "slots" within an event-level template of expected actions and indicators. Automatic Performance Assessment (APA) software then compared the observed behavior of the participants with the actions that would be expected by a qualified performer (Carolan, Bilazarian, and Nguyen, 2005). The APA system recorded differences between the observed and expected values for each measurement "slot" in the template, and assigned a numeric score accordingly. The DDSBE system also recorded the trainees' audio communications, and automatically captured screen shots of the trainees' tactical displays at ten second intervals. Instructors also could request additional screen captures via the VCAT tool.

Table 1 presents the 28 performance measurement data items collected by the DDSBE system for each air-to-air engagement. The measures are listed in chronological order and grouped by engagement phase. Eleven of the measures were collected by the automated data collection system that recorded the ACO's and CICO's keystrokes and mouse clicks and the Sweep Lead's control stick movements and button presses.

The remaining 17 measures were manually collected by the instructors using the VCAT device. Observation

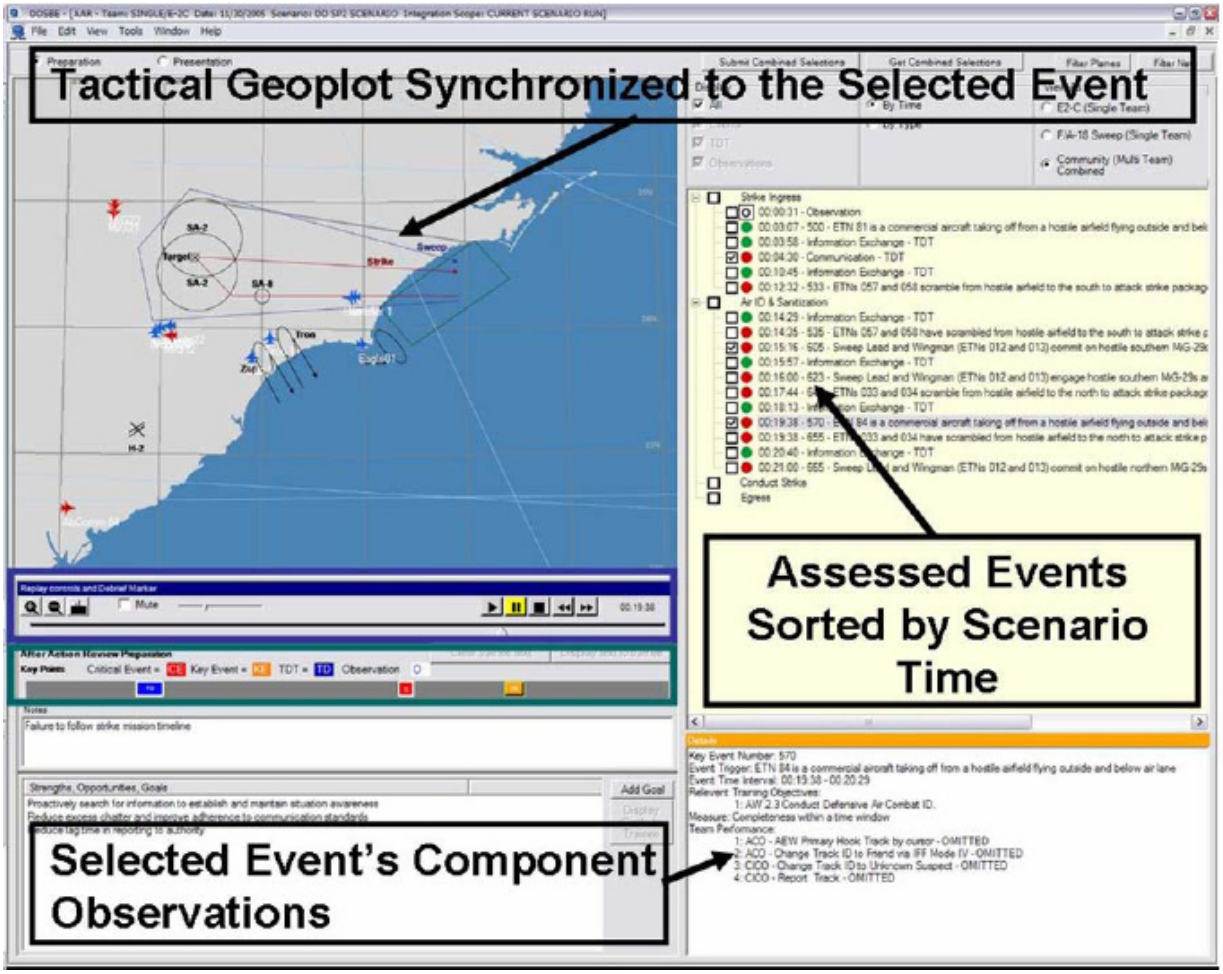


Figure 1. DDSBE AAR Interface

scores were used to compute four event-level scores that were aggregated with scores on other relevant events to compute scores for the scenario's training objectives. The individual observation scores also were used to compute mastery scores on individual, team-level, and mission-level competency scales. At the end of the scenario the collected performance data, the track position data, and the accompanying audio and visual recordings were compiled by Assessment Integration software and presented to the instructors for preparation of a debrief. Figure 1 presents the interface of the DDSBE AAR preparation and delivery tool.

The DDSBE AAR tool (Freeman, Salter, & Hoch, 2004) was designed to present the performance data aggregated in chronological order at the event level and by scenario training objectives. Individual events were labeled with "traffic light" symbols of green, yellow, or red to indicate the performance score assigned to the trainees on the event. The red and yellow symbols indicated events in which trainees had performed at a less than acceptable level on one or more tasks or steps within the event. The instructors could "drill down" into an event to identify the specific performer (e.g., CICO) and the performance details (e.g., a missed report) that resulted in the team's score on an event.

The AAR tool also permitted instructors to assess performance in the context of the overall strike mission timeline. When an instructor selected an event from the list on the right of the screen, the geographic display to the left automatically presented the location and heading of all tracks at that moment in the scenario. An instructor also could replay the audio communications and the trainees' tactical displays during the event. Thus, an instructor could present both the assessment of the event and the evidence supporting that assessment in the context of the overall situation.

F/A-18 AIRCREW TRAINING RESEARCH

The DDSBE project also developed F/A-18 pilot performance data using virtual and constructive entity position-derived data collected from the distributed network. However, limitations in the simulation environment and project priorities reduced the number that could be tested in the experimental runs described earlier. Therefore, a second research project was initiated through a Cooperative Research and Development Agreement (CRADA) between the Naval Air Warfare Center Training Systems Division (NAWCTSD) and The Boeing Company, Training System & Services (TSS). This complementary project focused on integrating and presenting automated measures of F/A-18 aircrew performance in order to identify strengths and weaknesses in the technologies and provide recommendations for improving the reliability and validity of automated assessment.

The performance assessment test bed was implemented by Boeing TSS and consisted of two F/A-18 unclassified simulators developed by Boeing, a Big TacTM air threat generator, an Instructor Operator Station (IOS), and automated data collection, analysis, and visualization software. Standard Distributive Interactive Simulation (DIS) network data and non-standard (e.g., button presses, instrument readings) simulation network data were logged with a DIS data logger.

Additional performance data collection was conducted by Alion Science and Technology, MA&D Operation, which was developing a Human-Centered Performance Assessment Tool (HCPAT) under a Small Business Innovative Research Phase II project. The HCPAT research project had developed automated and semi-automated performance measurements to evaluate the F/A-18 aircrew during the engagement and merge phases described earlier. Alion integrated the HCPAT with the Boeing test bed to implement automated and semi-automated metrics for testing. DIS communication middleware was developed as a plug-in to HCPAT to allow the software to observe the network traffic for relevant simulation entity state data; an air combat domain plug-in was created to specify the relevant objects in the performance assessment environment.

The F/A-18 stations were used by the Strike Lead and Wingman roles, and the IOS was used to support an E-2C role-player. The purpose of the E-2C role was to support the information exchanges that are part of the engagement and merge phases, but was not a focus of the performance assessment. The missions were geographically located in the vicinity of Elmendorf United States Air Force Base, Alaska.

Similar to the DDSBE approach, the EBAT methodology was used to fine-tune the scenario and guide the automated and semi-automated measures. A task analysis by Brobst, Geis, and Brown (1999) that organized the performance measures by the F/A-18 aircrew performance elements, air crew skill, and mission phase was leveraged as the basis for organizing the metrics into competencies. A list of scenario events expected during each mission phase was created, and expected tasks and actions were linked to each event. Measures and performance standards were created for a sample of the event tasks and selected for implementation based on mission requirements input from the Subject Matter Experts (SMEs) and the simulators' capabilities. Metrics were designed to generate automatically or with observer input depending on the data available from the flight simulators.

A secondary objective was to identify technical data requirements for constructing specific F/A-18 performance

elements. Metrics that only required DIS data could be implemented on a standard DIS network. However, access to non-standard DIS data required software modification. In this experiment, non-standard DIS data was obtained through a Protocol Data Unit (PDU).

Four post- Fleet Replacement Squadron-level air-to-air scenarios, each successively more difficult, were scripted by an F/A-18 Subject Matter Expert (SME). The scenarios involved two F/A-18 pilots (Strike Lead and Escort Lead) and an E-2C role-player (ACO). The experiment was designed to analyze the reliability and validity of the metrics across two performance levels within scenarios of differing difficulty. The four scenarios were each performed by the SMEs three times; once to standard, and twice not to standard.

Maintain Mission Timeline
• Time and distance off at waypoints
Weapon Launch
• Range at missile launch
• Clear avenue of fire
• Tactical advantage – Relative speed and altitude
• Acceptable launch region
• Crank maneuver
Defensive Maneuvers
• Within E-Pole range/orientation to threat
• Escape maneuver executed
• Maximum G-force attained
• Time to achieve escape range and heading
Maintain Mutual Support
• Outside mutual support range or altitude
• Outside contract speed and altitude value ranges

Table 2. Automated F/A-18 Aircrew Performance Measures

In the non-standard performance conditions, the F/A-18 pilots deliberately exhibited pre-specified behaviors to test the metric's ability to accurately report the greater variability in performance. Each mission was designed to affect performance on a specific training objective. The missions followed the generic fighter engagement timeline described in the background section. The timeline was adapted to the experimental mission timeline and varied by the complexity of the threat fighters' performance in the four conditions.

Automated air combat measures were developed based on existing air to air combat algorithms developed for the Navy DDSBE project (Carolan, Bilazarian and Nguyen, 2005), analyses and measures developed for the Air Force (Portrey, Schreiber, & Bennett, 2005) and new algorithms developed for the Boeing aircrew training research environment. The approach was to capture performance relevant data and use the data in metrics to evaluate warfighter performance with respect to higher level training objectives, such as aircrew tasks, mission phases, and underlying competencies. Some of these measures are considered first approximations since not all the data to accurately compute the measure was available. Table 2 presents automated F/A 18 aircrew performance measures developed and tested.

Observer-based measures were constructed for most of the task areas. These were focused on communications - the completeness, timeliness and format of voice reports, adherence to task procedures, and tactical decision making. In addition teamwork measures were also available to the evaluator. These were not event specific measures but provided the opportunity to assess specific aspects of team work observed throughout the scenario.

During the exercise an evaluator using a networked tablet style computer with the HCPAT software observed performance, selected events to assess and entered assessment data. The assessed events were displayed on an event log. The evaluator had the option of entering events and completing the assessments at a later time. The single evaluator assessed between 10 and 20 events during each exercise run. These items included the timeliness and completeness of voice communications, and the quality of tactical decisions.

The automated assessment module monitored the scenario entity state data through the DIS connection, detected performance events, and triggered measures. The performance events and measures were recorded in the event log and made available to the evaluator. An additional alert feature indicating that an event of interest had occurred was still in development and not available during the test.

The automated performance measures were designed to record deviations from expected performance standards, as in Outside Of Mutual Support Range, or a value to be compared against performance standards such as Within E-Pole Range. Automated measures can be event specific measures or global measures monitored as appropriate throughout the scenario. Global measures consisted of detecting and flagging observed deviations from expected performance criteria.

In addition to fully automated measures, which required no human intervention, semi-automated measures were employed to support the observer assessment process. One example is the automated calculation of time between events, where one event is an observer selected voice report. Another measure is the range between entities when a particular event is triggered. Since these are

based on evaluator response time, they provide estimates to support the evaluator's assessment.

This is the initial step of the assessment process. The deviations are recoded in the event log and linked to higher level measure categories through the structure of the event tree or through predefined analysis groups. The software supports a number of approaches for using this performance data for assessment and feedback. The first approach uses the automated performance measures to support the evaluator in making assessments. For many of these dynamic measures the assessment can be very context dependent. Flagging potential problem areas and providing the evaluator with performance evidence, behavior anchors, and a rating instrument, allows the evaluator to make the assessment based on observations, context and performance evidence. Simple examples include Maintaining Mutual Support, staying with contracted speed and altitude ranges. We found multiple departures from mutual support range under the 'good' performance conditions. Many were small departures, others were larger, such as, to investigate a potential threat. The evaluator reviews the performance data and makes the judgment on how to assess.

A second approach is to build in automated assessment algorithms that assign a value to a performance instance or set of instances based on predefined standards and context information. Some of these assessments are built into the measure, such as, a simple pass or fail for clear avenue of fire. Others require triggers to turn measures on and off. In addition, other standards change depending on whether they are performed pre- or post-commit. Others require a more detailed situation assessment and expected performance model, such as assessing targeting decisions.

Real-Time and Post-Event Analysis and Presentation

The Evaluator is an automated analysis tool prototype that can be used to create metrics in near real-time during the performance of a training mission and/or on completion of

Figure 2. Real-time ResultsViewer Display

a training mission. We used the post-event evaluation approach in order to create a quantitative, summative value for the measures we collected during the experimental scenario runs. For example, the Maintain Mutual Support metric returns the average range (in nautical miles) between the aircraft over a period of time. If the average distance is within an acceptable maximum range the Maintain Mutual Support value can be further qualified as a "pass." These results can be displayed as "passed" (green) or "fail" (red), based on the raw metric result. Metrics coded yellow, (e.g., Shot Kinematics) involved two quantifiable variables – in this case, the altitude and speed of the aircraft at the time the shot was made. If only one parameter was within standard, the metric evaluated as "partial pass," and displayed with a yellow symbol. The post-event analysis method was used to verify that the SME's performance was assessed as intended.

A complete analysis of the data we collected is still under review. However, initial findings from the experiment enabled us to identify critical weaknesses in the simulation and assessment system that pointed to needed improvements in technologies. Although SMEs had performed to pre-scripted actions, the post-event analysis indicated their actual performance on the scenarios, in many cases, did not match expected performance on the measures. An in-depth analysis of the raw metric data and post-event discussions with SMEs provided valuable insights on the major causes of the inconsistencies in the assessment system results as described in the following.

Mismatch between expected performance and simulation test bed design. Although the performance metrics we developed were specified according to real world F/A-18 pilot behaviors, the simulation test bed lacked some critical functionality in order to be implemented in an unclassified environment that would have allowed the SMEs to perform to expectations. We understood in advance that some of the SME actions would be "artificial" compared to real world behaviors, and as it turned out, the assessment results enabled us to identify this problem.

Task complexities. The parameters used for evaluating performance may have been too constrictive given the complex nature of some of the pilot's tasks. The SME's review of post-event analyses enabled us to understand the extent of the complexities of the performance elements that the metrics were assessing as well as the situation-dependent nature of the metrics.

Accuracy of performance measures algorithm. In some cases the performance measures algorithms did not accurately evaluate the task. The process of evaluating the data and talking to the SMEs enabled these metrics to be refined.

Figure 2 presents a snapshot of sample performance data presented in the realtime ResultsViewer display. It is a prototype data visualization tool that is used to display the near, real-time metrics during the performance of a training mission or during a mission playback, such as during an AAR. The real-time ResultsViewer approach is to provide the instructor with a graphical display of the metric as it evaluates data in near real-time. This approach can be used during the performance of the training exercise or the ResultsViewer can be played back in synchronization with a mission playback during the debrief session. These displays can be used to alert the instructor to a particular situation that may not be detectable through human observation or due to the complexity of the many events that the instructor must simultaneously observe.

The advantages of an integrated assessment approach is it can provide different automated performance data to training evaluators and training participants at different times during or post exercise to support ongoing assessment, diagnosis, and performance feedback needs at different levels of analysis. With a focus on providing formal assessment (ratings) for AAR, one HCPAT product is a drill down assessment report implemented as a set of PowerPoint slides. The assessment report displays the color coded ratings and associated comments at each level down to the performance instances. The AAR leader can start at the highest level; for example, the mission phase, or Mission Essential Task level, and then drill down to specific performance instances in the context of the overall scenario situation.

CONCLUSION

Both the DDSBE and F/A-18 Aircrew Training Research systems provided an opportunity to test and evaluate different approaches to collecting, analyzing, and presenting performance data regarding team and collective performance in a distributed simulation training environment. This type of experiment was critical to identifying the complexities, strengths, and weaknesses of automating assessment of team performance. The following guidelines are based on the results and feedback received during the various experiments.

1. **Use the EBAT approach for scenario and performance measurement design:** The EBAT approach involves the development of performance measures and data collection requirements during the scenario design process. Therefore, human observation requirements are pre-defined, which will result in minimized workload and simplified data collection processes. This will serve to improve the reliability and validity of the data collected and subsequent assessment. Additionally, the EBAT methodology reduces the tendency to

collect data on "everything." Experience has shown that this method does require clear segregation of events that do not influence each other, and occur in the order expected. In order to prevent a reduction in the realism of the scenarios due to these constraints, and to allow for assessment when performers react to events, it is important to develop flexible event-based metrics that can adapt to the context of the scenario in real time (e.g., Biddle, Perrin, Dargue, Lunsford, Pike, and Marvin, 2006).

2. **Concentrate performance assessment on known, verified (or verifiable) relationships between observed behavior and likely gaps in certain competencies:** Focusing the assessment on known relationships between specific behaviors and gaps in competencies facilitates the diagnosis of root cause. These relationships are now found largely in the experience of SMEs and remain to be captured by training practitioners. Consequently, this diagnostic process needs to be supplemented with human observation during the event to verify that root cause diagnosis is accurate and not due to an unforeseen event or training system failure.

3. **Focus on specific events vice general observations (i.e., "You need to improve your communications!") during debriefs:** The use of specific events from the training scenario to discuss an instructional point will improve instructional benefits by providing feedback in context of a specific event. So that expert instructors do not feel excessively controlled by the focus on specific events and specific observations, the post-event automated results, in conjunction with post-event semi-automated results, can be used to provide global observations and evaluations, as long as the instructor can then point to specific events in the scenario.

4. **Focus analysis on processes as well as outcomes:** The integration of process and outcome measure assists in providing understanding of how team and individual behaviors contribute to event outcomes.
5. **Use graphical presentation of performance measures updated in near real-time:** Real-time visualization of performance can be used to assist or alert the instructor in diagnosing trainee performance problems and providing real-time feedback or scenario modification. Additionally, the real-time assessment information provides instructors with detailed information regarding student performance that may not be obtained through human monitoring or objective analysis. Real-time visualizations do not provide an overall report on the metric so it should be used in combination with post-event metric results.

6. **Balance real-time and post-event automated performance assessment and scoring:** A summative, post-event metric provides a quantitative value to provide meaning regarding a "pass" or "fail" evaluation. Real-time ratings may be based on incomplete or premature interpretations of events. The results of both processes need to be considered in conjunction with each other to produce the most accurate and useful feedback to the trainees.

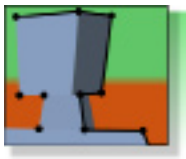
These guidelines are by no means fully-conclusive, and the authors recommend that research continue in this area to enable greater reliability and validity in automating the test and evaluation of training effectiveness. In many cases, the guidelines are more cautionary than prescriptive, which also argues for more thought and testing in this area. The challenge is to integrate and analyze objective performance data from simulation environments so that it is useful for assessment, diagnosis and feedback. This means analyzing both the capabilities of the simulation to support the performance of the tasks being trained or assessed, and the degree to which the data produced in the simulation reflects the trainees' competence to perform those tasks in the real world. It also underscores the importance of empirically testing and validating all aspects of the performance measurement system.

MEMORIUM

We dedicate this paper to the memory of Paul Radtke who passed away during the time it was completed. In his 18 years as a top notch Navy scientist, Paul worked hard to achieve many successes in transitioning scientific products to the research community, the schoolhouses, the operational Navy, and to our joint and coalition partners. He was a great friend, collaborator, and a mentor to all of us, always finding ways to make our work together both effective and fun. We will miss him very much.

ACKNOWLEDGEMENTS

The authors would like to extend thanks to Hugh Carroll (BGI), Steve Dix (Boeing), David Fries (Boeing), Mike McCleod (Thunderbolt), Jeff Miller (Alion), Richard Plumlee (NAWCTSD), LCDR Chris Provan (NAWCTSD), Erick Weber (Alion), and Jake Wigglesworth (Boeing) for their efforts in planning the scenario, specifying performance measures, and participating as role-players during DDSBE and Augmented DDSBE experiments.



Integrating and Presenting Performance Information in Simulation-Based Air Warfare Scenarios

REFERENCES

Biddle, E., Perrin, B., Dargue, B., Lunsford, J., Pike, W.Y., & Marvin, D. (2006). Performance-based advancement using SCORM 2004. In the Proceedings of the 2006 Interservice/Industry Simulation, Training, & Education Conference [CD-ROM]. Orlando, FL.

Brobst, W.D., Geis, L.A., & Brown, A.C. (1999). NSAWC aircrew training study: Methodology and analysis (Report No. CRM 98-171). Alexandria, VA: Center for Naval Analyses.

Carolan, T. F., Bilazarian, P., & Nguyen L. (2005). Automated individual, team, and multi-team performance assessment to support debriefing distributed simulation based exercises (DDSBE). In the Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting [CD-ROM], Orlando, FL.

Fowlkes, J., Dwyer, D. J., Oser, R. L., & Salas, E. (1998). Event-based approach to training (EBAT). International Journal of Aviation Psychology, 8, 209-221.

Freeman, J., Salter, W.J., & Hoch, S. (2004). The users and functions of debriefing in distributed, simulation-based team training. In the Proceedings of the 48th Annual Conference of the Human Factors and Ergonomics Society [CD-ROM], New Orleans, LA.

Glassburn, J. (2006, April). U.S., coalition forces conduct at-sea training without leaving the pier. Navy Newstand Online. Retrieved June 11,2007, from, http://www.news.navy.mil/ search/display.asp?story_id=22943

Jean, G. (2006, September). Navy's virtual training exercises expanding in realism and scope. National Defense Online. Retrieved June 11, 2007, from http://www.nationaldefensemagazine.org/issues/2006/September/NavyVirtual.htm

Johnston, J. H., Radtke, P. H., Van Duyne, L., Stretton, M., Freeman, J., & Bilazarian, P. (2004). Team training in distributed simulation-based exercises. In the Proceedings of the 48th Annual Conference of the Human Factors and Ergonomics Society [CD-ROM], New Orleans, LA.

Neville, K., Fowlkes, J., Milham, L., Merket, D. C., Bergondy, M. L., Walwanis, M., & Strini, T. (2001). Training team integration in a large, distributed, tactical team: A cognitive approach. Proceedings of the 23rd Annual Interservice/Industry Training, Simulation and Education Conference (pp.1035-10), Orlando, FL,

Portrey, A. M., Schreiber, B.T., & Bennett, W., Jr. (2005). The pairwise escape G-metric: A measure for air combat maneuvering performance. In the Proceedings of the 2005 Winter Simulation Conference (1101-08), Orlando, FL.

ABOUT THE AUTHORS

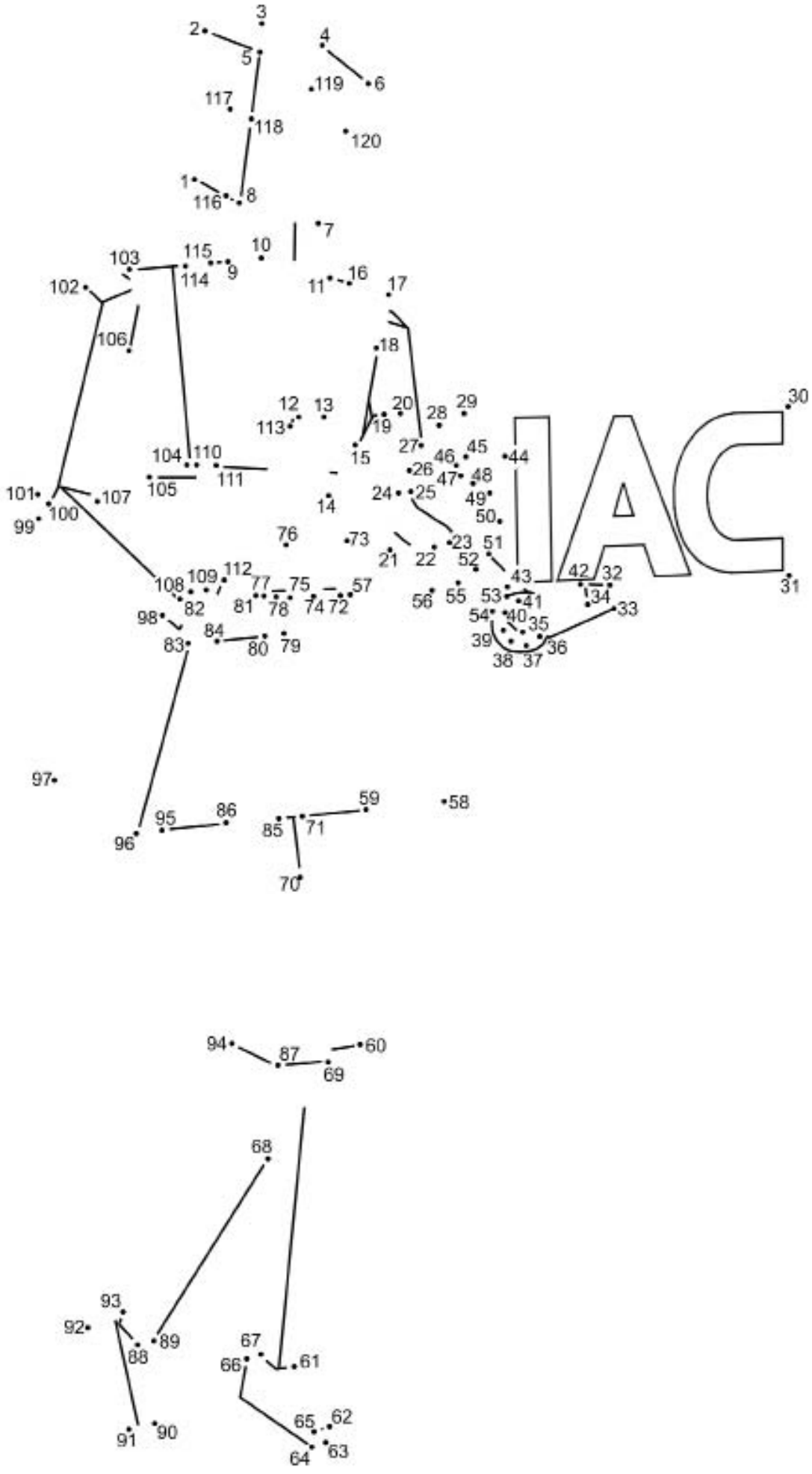
Paul Radtke (*paul.radtke@navy.mil***)** is a Research Psychologist with the Naval Air Warfare Center Training Systems Division, Orlando, FL. He holds a BA in Political Science from Western Illinois University and completed the MA in Political Science at Northern Illinois University. Before coming to NAWCTSD in 1994, he served as a Personnel Research Psychologist at the Navy Personnel Research and Development Center in San Diego, CA.

Joan Johnston, Ph.D. (*joan.johnston@navy.mil***)** is a Senior Research Psychologist and a NAVAIR Associate Fellow at Naval Air Warfare Center Training Systems Division, Orlando, FL. She is responsible for managing basic, applied, advanced technology development, and prototype training research. Dr. Johnston's technical research areas are tactical decision making under stress, team performance and team training technologies, and distributed simulation-based training. She received her M.A. and Ph.D. in Industrial and Organizational Psychology from the University of South Florida.

Elizabeth Biddle, Ph.D. (*elizabeth.m.biddle@boeing.com*) is a Manager with Boeing Training Systems & Services. She has served as Principal Investigator for advanced instructional and training research and development projects. Dr. Biddle earned a Ph.D. in Industrial Engineering and Management Systems from the University of Central Florida in 2001.

Tom Carolan, Ph.D. (*tcarolan@alionscience.com)* is a Program Manager with Alion Science and Technology, MA&D Operation. He received his Ph.D. in Experimental Psychology from the University of Connecticut. He has been involved in research and development related to training systems, performance measurement, and human performance modeling for the past 19 years.

This paper was adopted from I/ITSEC 2007 and is reprinted with permission.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120