

EVALUATING THE EFFECTIVENESS OF A JOINT COGNITIVE SYSTEM: METRICS, TECHNIQUES, AND FRAMEWORKS

Scott S. Potter
ManTech International
Corp.
Pittsburgh, PA

David D. Woods
The Ohio State
University
Columbus, OH

Emilie M. Roth
Roth Cognitive
Engineering
Brookline, MA

Jennifer Fowlkes
CHI Systems
Orlando, FL

Robert R. Hoffman
Institute for Human and
Machine Cognition
Pensacola, FL

An implication of Cognitive Systems Engineering is that joint cognitive systems (JCS; also known as complex socio-technical systems) need to be evaluated for its effectiveness in performing the complex cognitive work requirements. This requires using measures that go well beyond “typical” performance metrics such as the number of subtask goals achieved per person per unit of time and the corresponding simple baseline comparisons or workload assessment metrics. This JCS perspective implies that the system must be designed and evaluated from the perspective of the shift in role of the human supervisor. This imposes new types of requirements on the human operator. Previous research in CSE and our own experience has lead us to identify a set of generic JCS support requirements that apply to cognitive work by any cognitive agent or any set of cognitive agents, including teams of people and machine agents. Metrics will have to reflect such phenomena as “teamwork” or “resilience” of a JCS. This places new burdens on evaluation techniques and frameworks, since metrics should be generated from a principled approach and based on fundamental principles of interest to the designers of the JCS. An implication of the JCS perspective is that complex and cognitive systems need to be evaluated for *usability*, *usefulness*, and *understandability*; each of which goes well beyond raw performance. However, conceptually-grounded evaluation frameworks, corresponding operational techniques, and corresponding measures for these are limited. Therefore, in order to advance the state of the field, we have gathered a set of researchers and practitioners to present recent evaluation work to stimulate discussion.

PANEL SUMMARY

We define a Joint Cognitive System (JCS) as the combination of human problem solver and automation/technologies which must act as co-agents to achieve goals and objectives in a complex work domain (cf. Hollnagel and Woods, 2005 for their cyclic model of the function of Joint Cognitive Systems). A key point to note is that the cycle emphasizes that current decision making builds on previous decisions, and anticipates future decisions. The humans and machines must work as a coordinated team to be successful in the variety of complex work domains that are being supported.

This JCS starting point implies several things. First, the JCS consists of the entire set of humans, technology, and automation systems operating as one team. Second, this system as a whole must be sensitive to the context in which it is currently operating. Lessons learned from several domains indicate that only the humans in the JCS have demonstrated the ability to successfully fill this context awareness/maintenance role. Third, any change in the level of autonomy affects the entire JCS and changes the requirements of the automation with respect

to interacting with the human. Fourth, decision-making must be considered from the JCS perspective (regardless of the agent making the decisions).

Most automation is developed with the purpose of replacing human involvement to both reduce the chance for human error and reduce human workload. This assumes that the new automation can be substituted for human action without any larger impact on the overall system. However, investigations of the impact of new technology have shown that this assumption is not tenable. Sarter, Woods, and Billings (1997) call this the “substitution myth”. More typically, from a JCS perspective, introduction of new automation has shifted the human role to one of monitor, exception handler, and manager of automated resources leading to a new class of “human error” types of problems. These problems are due, for the most part, to breakdowns in the interaction between the human operators and the automation; even highly automated systems must be able to be managed by the human operator and, therefore require a significant degree of decision support capabilities. Because of the substitution myth, however, these requirements are typically not supported by most

systems. Rather, they are designed to be precise and powerful agents, but are not equipped with decision support capabilities necessary to be co-agents rather than substitutes, not given comprehensive access to the outside world or knowledge about the tasks in which it is engaged.

The Joint Cognitive System philosophy has a significant impact on the design and evaluation requirements for automation. An extensive history of problems and accidents with automated systems has consistently shown that they are largely the result of “technology centered” designs that did not consider this need for supporting communication and cooperation between human and machine agents. To counteract and prevent the reoccurrence of such difficulties, “human centered” automation must be developed. Its objective is to support, not supplant or replace the human operator. The primary focus becomes how to make automated systems team players.

This JCS perspective implies several things. First, the system must be designed and evaluated from the perspective of the shift in role of the human supervisor. This imposes new types of requirements on the human operator. Previous research in CSE and our own experience has led us to identify a set of generic JCS support requirements that apply to cognitive work by any cognitive agent or any set of cognitive agents, including teams of people and machine agents (Billings and Woods, 1994; Dekker and Woods, 1999; Christoffersen and Woods, 2002; Tittle, Elm, and Potter, 2005). These include:

- **Observability** – the ability to form insights into a process (either a process in the work domain or in the automation), based on feedback received. Observability overcomes the ‘keyhole’ effect and allows the practitioner to see sequences and evolution over time, future activities and contingencies, and the patterns and relationships in a process.
- **Directability** – the ability to direct/redirect resources, activities, and priorities as situations change and escalate. Directability allows the practitioner to effectively control the processes in response to (or in anticipation of) changes in the environment.
- **Teamwork with agents** – the ability to coordinate and synchronize activity across agents. This defines the type of coordination (e.g., seeding, reminding, critiquing) between agents. Teamwork with agents allows the

practitioner to effectively re-direct agent resources as situations change.

- **Directed attention** – the ability to re-orient focus in a changing world. This includes issues like tracking others’ focus of attention and the ease with which they are interrupted. Directed attention allows the human-system team to work in a coordinated manner, resulting in increased effectiveness.
- **Resilience** – the ability to anticipate and adapt to surprise and error. This includes issues such as failure-sensitive strategies, exploring outside the current boundaries or priorities, overcoming the brittleness of automation, and maintaining peripheral awareness to maintain flexibility.

These requirements define success for the JCS and as such define evaluation criteria that are essential for an effective JCS. These requirements take situation awareness from an ill-defined concept to tangible, measurable concept. In addition, they define a whole new class of decision-making that must be evaluated in order to ensure an effective JCS team. These requirements specify what it means for a complex cognitive system to be “good” in terms of human-centering and human-systems integration.

However, most new decision aids are typically evaluated in two ways, one being performance measurement and the other being satisficing. In the satisficing procedure, in the final spiral prototype is presented to domain practitioners, who work with it for some short time, and then provide an evaluation, which is often a questionnaire-like evaluation of what they liked and did not like. Based on that evaluation, a final re-prototype is created and then taken to the field or simulated-field testing stage. This procedure is likely to involve task demand characteristics, and complications due to buy-in and cognitive dissonance. Furthermore, it does not involve any structured empirical assessment of strengths and limitations. In the cognitive systems engineering community, spiral modeling is notorious for actually leading to the creation of systems that are user-hostile.

In the performance measurement procedure, counts are made of aspects of “raw” performance. These are typically things that are easily measured (e.g., number of tasks executed per unit time, number of sub-goals achieved per unit time, time-to-goal, etc.). Thus, for instance, for a new military command post workstation, measures might be number of enemy tanks destroyed per period of watch. This procedure is likely to involve standard scenarios to facilitate comparisons between

systems, but does not address the resilience of the JCS in the face of novel situations outside the boundaries.

An implication of the Laws of Cognitive Work is that complex and cognitive systems (also known as complex socio-technical systems) need to be evaluated for *usability, usefulness, and understandability*; each of which goes well beyond raw performance. The primary JCS support requirements enable us to specify the high-level mental and coordinative functions that complex cognitive systems must support. However, conceptually-grounded evaluation frameworks, corresponding operational techniques, and corresponding measures for these are limited. The JCS support requirements enable us to predict with some confidence the bad things that will happen using a new system. When and if those predictions come true, attention focuses on fixes and new procurements. But predicting on the basis of cautionary tales, however lawful, is not enough. Things have to be measured.

A survey of the literatures of human factors and cognitive systems engineering suggests that many other things need to be measured, most of which go beyond measures of raw performance because it is these things that either underlie or undermine performance. Some measures exist – measures of mental workload, for instance. This includes self-rating scales such as NASA-TLX, and experimental procedures such as dual-task interference.

The extent to which such measures can be adopted for use in the on-going evaluation of complex and cognitive systems is not yet clear. Consider, for example, “avoidances.” This has not been considered important enough to actually baseline and measure in systems evaluation, but experience shows that they are significant considerations when new systems are fielded. Observers of team activities, including those of the military, invariably notice how workers have created kluges and work-arounds, and are burdened with make-work. Further, it is likely that in the immediate future, new technologies that will be fielded will *still* have features that require activities that should be avoided. It is recognition of “avoidances” that serves to identify leverage points for how the work might be improved. But apart from observation by humans (who have an eye trained to notice these things), we have no formal, let alone robust, measures or measurement procedures. Only when we do, can we be judicious in making the really tough decisions about how to address the deficiencies identified by the evaluations.

Therefore, in order to advance the state of the field, we have gathered a set of researchers and practitioners to present recent evaluation work to stimulate discussion.

Each will present the essential issues from their work and then there will be time for discussion at the end of the session.

Panelist #1: Dr. David D. Woods Measuring Adaptive Capacity and Agent-Environment Fit

I will explore how Cognitive Engineering can begin to assess the adaptive capacity of joint systems by analyzing three examples of new metrics:

- fractal path analysis as a means to assess remote presence developed by M. Voshell and F. Phillips;
- the modified Unit Marking Procedure that captures the events practitioners find interesting as a means to assess direct perception as developed by K. Christoffersen, G. Blike, and D. Woods;
- re-orienting cost as a means to assess control of attention in multi-task situations as developed by G.-D. Tuzar and D. Woods.

Interestingly, each of these metrics is sensitive to agent-environment interactions as is important in ecological approaches to human-machine systems (Flach et al., 1995).

Based on understanding common properties of these three metrics, I propose a set of targets for measuring adaptive capacity of joint cognitive systems. Adaptive capacity can be thought of as three sets of contrasting states following a change in the organization or technology of work. When one assesses the impact of a change they examine evidence of how the people in the system adapt. This adaptation takes three forms:

- how much does the change create complexities to be worked around versus how much does the change stimulate incremental improvement?
- how much does the change broaden the field of awareness in time and space or how much does it narrow the field of awareness?
- how does the change lead people to tune current work practices versus how does the change lead people to innovate new strategies that exploit new capabilities.

Panelist #2: Dr. Emilie M. Roth A Work-Centered Approach to Evaluation

As Woods and his colleagues (Woods, 1998; Potter, Roth, Woods and Elm, 2000; Woods and Dekker, 2002) have argued, new support technologies should be regarded as hypotheses about what constitutes effective

support. A key premise of work-centered design and evaluation is the importance of incorporating work-centered ‘looks’ – up to and beyond the point where the system is fielded – to help insure that a support system will be successful in the intended work context (Eggleston, 2003; Eggleston, Roth and Scott, 2003; Scott, et al., 2005).

Work-centered evaluations simultaneously address formative and summative questions. From a summative perspective the aim is to assess whether the proposed design concepts (e.g., as embodied in a prototype) have the positive effects predicted by the design developers. From a formative perspective the aim is to uncover additional demands and unanticipated requirements. This includes probing for the boundaries of effectiveness and breakdown conditions where the system no longer provides effective support. The results can then be used to propel further design and expand the cognitive engineering theoretical base.

Panelist #3: Dr. Jennifer Fowlkes (with Kelly Neville, Robert Hoffman, & Jerry Owens)
Checks and Measures to Support Design Teams in Building Complex and Cognitive Systems

In recent years, traditional engineering models have been criticized from the standpoint of their suitability for developing large, integrated systems. Criticisms and concerns have been expressed within the systems development and procurement community (e.g., Nusiebh, 2001) and within the human factors community based on the study of complex systems (e.g., Bar-Yam, 2003) and cognitive engineering (e.g., Hoffman & Elm, 2004). This contribution reports on a specific effort to develop a joint systems engineering method (JSEM) (Neville, Fowlkes, Hoffman, & Owens, 2006). One of the goals of the JSEM work was to characterize shortfalls of traditional models using methods such as semi-structured interviews with systems and software engineers, review of case studies, and literature synthesis and to suggest process improvements. The project team found that shortfalls in traditional engineering models can be understood in terms of at least four key themes: Coping with system complexity; managing changing requirements; supporting user needs; and, underlying the other themes, facilitating design team collaboration. The specific purpose of this contribution is to describe the four themes and then reference them in developing an organizing framework to identify checks and measures that can be used by design teams. The purpose of the checks and measures is to improve internal team coordination processes and to focus design teams on assessing the robustness of systems in meeting the

challenges inherent in today’s complex operating environments.

Panelist #4: Dr. Scott S. Potter
Evaluating the Net Decision-Making Effectiveness of the Joint Cognitive Team

New results from a unique type of evaluation technique – Decision-Centered Testing – will be discussed. DCT is conducted to evaluate whether or not the new joint cognitive system truly demonstrates increased decision-making effectiveness – both within and outside the boundaries of normalcy. DCT involves explicit design and analysis of tests based on the key decision making problems within the domain. The result is be an explicit test design describing the cognitive problem under test, the decision-making context that must be established, as well as the events that need to be included in the scenario. In DCT, test scenarios are developed to specify a progressive evolution of events that would be expected to be a difficult cognitive problem for the decision maker. This decision-centered approach to testing has proven effective in discovering fundamentally new ways for evaluating the net decision-making effectiveness of the joint human-technology decision-making team. The critical aspects of this technique are:

- Focusing the evaluation on an analytical model of cognitive demands of the work domain;
- Identifying the “edges” in the joint cognitive system for the particular focus of the evaluation;
- Defining scenarios explicitly based on this analytical basis in order to exercise the desired cognitive demands;
- Defining “cognitive pressure” to explicitly stress the edges and therefore assess the strength of the JCS.

In order to demonstrate the application of this technique, we will present its application to a decision-making micro-world. The results provided a powerful demonstration of the benefits of structuring the evaluation from a JCS perspective. Insights gained from this application are generalizable to other, more complex JCSs.

Panelist #5: Dr. Robert R. Hoffman
The Procurement Woes

Most system designers and human factors engineers have participated in projects that culminated in systems that were highly constrained by short-term cost considerations. In the procurement of information processing and intelligent technology for complex socio-

technical domains, the focus on short-term cost considerations at the expense of human-centering considerations always comes with a hefty price down the road, a price that weighs much more heavily on the shoulders of the end-users than on the shoulders of the technologists or project managers. Designer-centered design, whether conducted under the guise of the Spiral Model or the Waterfall Model, results in the sort of user-hostile technology that we see everywhere (e.g., VCR remote control devices) that frustrates people at home and at work. Negative and usually unanticipated consequences include bewilderment, road-blocking, clumsiness and make-work (need for kluges), displeasure, and automation surprise.

Worse, people tend to measure the things that are easy to measure, which often are the wrong things to measure. So, as the military commander uses a new tool, it might keep track of “number of tanks killed per hour.” Such convenient measures may *seem* to be direct reflections of the primary tasks or goals. They certainly leave much to be desired, specifically the things that are meaningful and therefore hard to measure, such as “support for accelerating the achievement of expertise,” or “feeling of immersion in the problem,” or “narrowing the actual work-true work gap.”

What concerns us is that in all of the documents about procurement, only occasionally does one see a reference to *guaranteeing on the basis of empirical evidence that the eventual technologies will help actual domain practitioners work on problems rather than forcing them to fight with the technology*. Even then, the requirements are stated as “physical/cognitive requirements” or “human performance effectiveness.” Rare is the explicit acknowledgement that systems should be proven to be both useable and useful; that they should motivate and not frustrate.

This presentation will discuss the implications of the principles of Human-Centered Computing for the procurement process, especially the methodological foundations for folding the exploration of the envisioned world into the reprototyping process, and help insure that systems are usable, useful, understandable, observable, and resilient. These “desirements” lead directly to a scheme for categorizing conceptual and operational definitions of measurables that would be valuable in the evaluation of complex cognitive systems.

REFERENCES

Christoffersen, K., Woods, D. D. and Blike, G. T. (in press). Discovering the Events Expert Practitioners Extract from Dynamic Data Streams: The mUMP Technique. *Cognition, Technology, and Work*.

- Voshell, M. G., Woods, D. D. and Phillips, F. (2005). Human-Robot Interaction: From Fieldwork to Simulation to Design. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*. 26-28 September, Orlando FL.
- Flach, J., Hancock, P., Caird, J., and Vicente, K. editors, *An Ecological Approach To Human Machine Systems I: A Global Perspective*, Erlbaum, 1995.
- Woods, D.D. and Hollnagel, E. (2006). *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*. Boca Raton FL: Taylor & Francis.
- Hollnagel, E., Woods, D.D. and Leveson, N., Eds. (2006). *Resilience Engineering: Concepts and Precepts*. Ashgate, Aldershot, UK.
- Eggleston, R. G. (2003) Work-Centered Design: A Cognitive Engineering Approach to System Design. *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting*.
- Eggleston, R. G., Roth, E. M. and Scott, R. (2003). A framework for work-centered product evaluation. *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting*.
- Potter, S. S., Roth, E. M., Woods, D. D. & Elm, W. (2000). Bootstrapping multiple converging cognitive task analysis techniques for system design. In J. M. Schraagen, S. F. Chipman & V. L. Shalin (Eds.) *Cognitive Task Analysis* (pp. 317-340). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Scott, R., Roth, E., Deutsch, S., Kuper, S., Schmidt, V., Stilson, M., and Wampler, J. (2005). Envisioning evolvable work-centered support systems: Empowering users to adapt their systems to changing world demands. *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*. (pp.244-248). Santa Monica, CA: Human Factors and Ergonomics Society.
- Woods, D. D. (1998). Designs are hypotheses about how artifacts shape cognition and collaboration. *Ergonomics*, 41, 168-173.
- Woods, D. and Dekker, S. (2002). Anticipating the effects of technological change: a new era of dynamics for human factors. *Theor. Issues in Ergon. Sci.* 1-11.
- Bar-Yam, Y. When systems engineering fails – Toward complex systems engineering. (2003). *IEEE Conference on Systems, Man & Cybernetic, October 5-8* (pp. 2021-2028).
- Hoffman, R. R., & Elm, W. C. (2004). *The Procurement Woes: Handcuffs on the Development of Intelligent Technologies* (Technical Report in review). Pensacola, FL: IHMC.
- Neville, K., Fowlkes, J.E., Hoffman, R.R., & Owens, J.M. (2006). *Joint Systems Engineering Method: Phase I Final Report* (CHI Systems Technical Report 060204.05016). Fort Washington, PA: CHI Systems.
- Nisebieh, B. (2001, March). Weaving together requirements and architectures. *IEEE Computer*, 34, 115-117.