Evaluating Air Campaign Plan Quality in Operational Settings[1]

Thomas E. Miller
Laura G. Militello
Jennifer K. Heaton

Klein Associates Inc.
582 E. Dayton-Yellow Springs Road
Fairborn, Ohio 45324-3987, USA
tmiller@klein.mhs.compuserve.com

Abstract

Artificial Intelligence methods can rapidly generate plans for complex situations. However, these plans may be rejected by air campaign planning experts, who judge dimensions such as robustness using operational rather than computational criteria. Our goal is to capture the tactical and strategic concerns of air campaign planners, and incorporate these into planning technology to filter out the unacceptable options and highlight preferred plans. We are in the process of employing a complete Cognitive Task Analysis (CTA), which uses a variety of knowledge elicitation techniques to fully identify the process of plan evaluation and factors underlying judgments of plan robustness. The CTA will form the foundation of an ACP Plan Critic. The Plan Critic will evaluate plans for plan quality and will highlight potential problem areas and vulnerable assumptions. This paper summarizes the current status of this work.

There is often a large gap between technological advances and operational feasibility. In the Artificial Intelligence (AI) planning domain, planning technology often approaches the problem via a process that does not match the way the targeted user thinks about and solves planning problems. In some cases, planning technologists and air campaign planners do not define certain concepts identically; what is robust to the planning technologist does not necessarily match robustness to the military planner. Subsequently, there is a considerable risk that current planning technologies will not be accepted in the field because they do not meet the decision-making needs of the commanders and planning staff. The following pages discuss the need for user-centered planning systems in the air campaign planning domain and describe a Decision-Centered Design approach to building systems, which will enable the generation of AI products that meet the operational needs of air campaign planners.

[1]This research is being supported by Rome Laboratory under contract F30602-95-C-0216.

The Air Campaign Planning Domain

The air campaign planning domain is in a state of transition. New approaches to planning air wars and advances in AI planning systems are converging on the air campaign planning domain. This set of conditions provides enormous opportunity to improve the efficiency of the planning process and ensure that plans of consistent quality are produced. These conditions, however, also carry the potential for misuse of technology and the development of systems which are unwieldy or lack credibility with the user community. The people who develop air campaign plans work under intense time pressure in high-stakes situations. They must develop a cohesive air campaign plan, but must use degraded and ambiguous information. Although this task requires a great deal of human judgment, decision making, and expertise, AI technology has much to contribute in streamlining this process. In order to apply AI technology appropriately so that the resulting system will successfully aid in the air campaign planning process, one must first understand the existing planning process.

Traditional approaches to campaign planning have taken a rollback approach; everything in the path of the advancing front is destroyed. Even in air campaign planning, the mentality toward targeting has been very linear in nature. After target lists are generated, strike packages are assembled for targets starting at the top of the list. Success in this tradition is defined in terms of absolute destruction, a "smoking hole" mentality.

New approaches to air campaign planning have been developed, such as Colonel John Warden's "Center of Gravity" approach. The term Center of Gravity describes the point where the enemy is most vulnerable and the point where an attack will have the best chance of being decisive (Warden 1988). Colonel Warden is credited with developing an early version of the Gulf War air campaign plan.

The Center of Gravity approach requires "strategy-to-task" planning and has led to the development of the Air Campaign Planning Tool (ACPT) under the initial direction of General Buster Glosson's organization. This approach is non-linear in nature and effectiveness is not necessarily measured in terms of absolute destruction, but in effective measures of destruction. For example, in World War II an air strategy was to bomb entire cities, whereas in the Gulf War, precise targets such as power station transformer yards were targeted with the intention to disrupt power for relatively short periods of time.

During Desert Storm, General Glosson focused on "deconstructing" Iraq. Rather than using a rollback strategy and attempting to achieve absolute destruction, specific targets were chosen which would meet the objectives of the US for the war. This approach leads to much more efficient and coordinated use of resources during implementation, but also calls for a more cognitively complex planning process.

This recent change in the approach to air campaign planning coupled with advances in the AI arena have created a situation which is ripe with opportunity to improve the planning process. New strategies toward air campaign planning have been developed and are being implemented in some areas of the Air Force community. Also, mature AI planning tools are being developed that can be applied in an operational environment. There is currently tremendous interest and excitement about the possibilities of building a mixed initiative air campaign planning system that follows the strategy-to-task approach and takes advantage of AI planning techniques where appropriate.

Researchers and developers in the planning subarea of artificial intelligence are interested in developing plan evaluation technologies as well. The advantages of such a system are numerous. For example, evaluation of an entire plan after it is complete can provide valuable feedback on the quality of the underlying knowledge used by the automated planner (so that, for example, a knowledge acquisition aid can improve the knowledge). Another advantage is that such mechanisms can permit the comparison of different planning systems. Also, evaluation mechanisms that can be called to quickly check partial plans can aid an automated planner in finding a good solution without a lot of backtracking. Automated methods of judging plan quality would aid the human planner in comparing and contrasting options and ensure that the human using the system produces the highest quality plan possible.

However, characteristics of a quality, robust air campaign plan are not well understood in this domain. The strategy-to-task approach has been documented and is used in some planning circles. However, the impact of the new strategy on plans and on the planning process has not been explored or documented, but potentially has tremendous implications for a Plan Critic. Further, the dynamics of the planning process are not well understood. These features must be understood if AI planners are to be successful in the operational community.

A skirmish with the Iraqis in early 1994 illustrates the environment in which a mixed initiative system will have to function. Personnel at Checkmate in the Pentagon were on

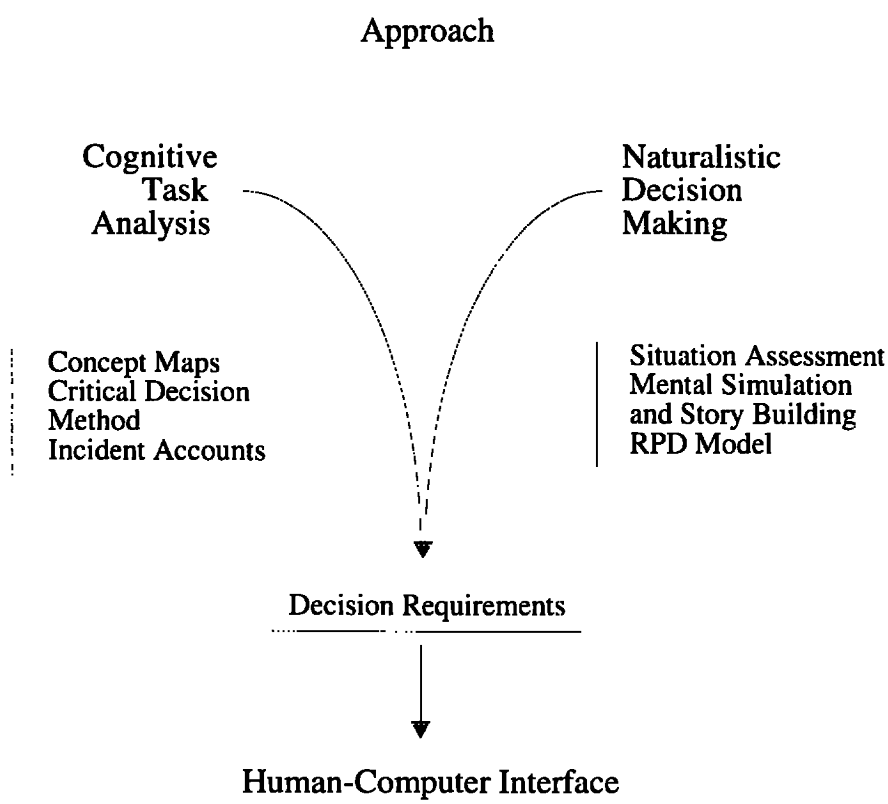

Figure 1. Decision-Centered Design approach.

24-hour work shifts for almost a full week. Events unfolded rapidly. The Iraqi military was building up in mass to the north of Kuwait, opposing a small allied force. Checkmate received a request to develop initial plans for how air power could be used in the situation. The initial plans were developed by a small group. This group of planners used a white board to outline initial assumptions and objectives. After the first few hours, they had developed a hand-written document that plan outlined a basic air power plan. Over the next few days, the plan was modified and refined as more information was gathered, as the Iraqi military changed their troop movements, and as objectives from the commanders changed.

Checkmate planners who took part in this process described a situation in which there were extreme time pressures because of the rapid buildup of Iraqi troops. One planner was only able to catch a 40-minute nap under his desk over a two-day period. With this kind of time pressure, high stakes, sleep deprivation, and required team effort, technological advances that do not address real concerns in the planning environment will not be used. Attempts were made to use new planning technologies during this incident, however, these technologies were not as responsive as needed in the rapidly changing planning environment.

Our goal is to address issues that will help with the development, transition, and acceptance of new technologies for developing quality plans in an operational setting. One aspect of plan quality that is under examination is that of robustness. Corresponding to our common sense understanding of this term, a computational definition of robustness would measure the ability of an implemented plan to achieve its goals, even in the face of changes to the constraints or conditions that were in effect when the plan was initially generated.

Decision-Centered Design Approach

Our approach is a Decision-Centered Design strategy (illustrated in Figure 1) which complements the technology-driven components of AI design planning technologies. This functional approach identifies the needs of air campaign planners based on the judgments and decisions they are required to make. Once identified, the primary decision requirements used to evaluate plans will form the foundation of the Plan Critic. The decision requirements are the specific judgments, diagnoses, and courses of action that planners generate when they evaluate plans. The decision requirements are identified using two primary inputs: the descriptive models of Naturalistic Decision Making, and the specific findings obtained through the Cognitive Task Analysis (CTA). Naturalistic decision models, such as the Recognition-Primed Decision model (Klein 1989), are helpful, but they are not sufficient since they are domain independent, descriptive models of decision making. The second input to decision requirements is the set of findings from the CTA. CTA uses different strategies to probe the way people perform tasks, e.g., the cues they rely on, the patterns they notice, the specific nature of the diagnoses and decisions they must make, and the reasons why these decisions are difficult. The output is a detailed understanding of how critical decisions specific to the domain are reached.

We are using three approaches in our CTA: the Critical Decision method (CDM), debriefings, and plan simulations. CDM interviews (Klein, Calderwood, & MacGregor 1989) use a set of cognitive probes that have been used in a great many studies and projects (e.g., Calderwood, Crandall, & Baynes 1988; Calderwood, Crandall, & Klein 1987; Crandall & Calderwood 1989; Crandall & Gamblian 1991; Crandall & Klein 1988; Klein 1989; Klein, Calderwood, & Clinton-Cirocco 1986; Klein & Thordsen 1988; Thordsen & Calderwood 1989; Thordsen et al. 1988; Thordsen, Klein, & Wolf 1990; Wolf et al. 1991). The cognitive probes used in CDM interviews are designed to help us learn more about tacit knowledge, such as the ability of firefighters to rapidly form a situation assessment and the ability of nurses to rapidly diagnose a preterm infant in the earliest stages of infection. The CDM has been adapted to the air campaign domain, and the probes are being used to identify which dimension of plan quality planners use to evaluate air campaign plans. In many cases domain experts are unable to describe the basis for their judgments, but after responding to the cognitive probes, they gain insights that are new, even to themselves, about how they form those judgments.

From our work to date, we have found that debriefing planners at Checkmate soon after a national crisis in which air campaign plans were developed is extremely fruitful. This data source is rich because the incident remains fresh in the planners' minds and they can relate their roles in the events, not just the "war stories" that evolve from a crisis. In the future, we would like to interview Checkmate planners as soon as they are available after a skirmish.

The third knowledge elicitation method used in the CTA is plan simulations (Thordsen, Militello, & Klein 1992; Crandall, Klein, Militello, & Wolf 1994). The interviewee is asked to evaluate a sample plan provided by the interviewer. The interviewees are asked to think aloud as they describe their evaluation. The interviewer then follows up with probes pertaining to the use of informational cues and specific strategies used to evaluate the plan.

We have analyzed preliminary interview data to develop an inventory of plan characteristics that influence plan quality and robustness. We also intend to document the planning process that experienced planners use. In a mixed initiative Plan Critic, it will be important to understand characteristics of good plans, how to evaluate those dimensions, how the system will need to keep the human in the loop, and where in the system either the human or the machine should be relied upon.

The findings from the CTA will enable us to design the

Plan Critic according to the operational needs of air campaign planners. The value of such a system to this community is that the system will automate many of the tracking and computational aspects of the task, so that the human user may spend more time and resources on the aspects of the task requiring human judgment, skills, and expertise. This will result in plans that are both more streamlined and of higher quality.

Preliminary Findings

From our CTA work to date, we have identified eight characteristics of good air campaign plans. These themes come primarily from interviews with planners at Checkmate. In the future, we will be conducting interviews with other planning communities such as the Air Force components of the Unified Command (e.g., PACAF) or significant planners from the Gulf War. These characteristics can be differentiated into characteristics of good plans and characteristics of planning done by experienced planners. These characteristics are summarized below:

Characteristics of good plans:

- Objectives must be clearly stated, measurable, and point toward an end state. Mission statements tend to be very broad, but specific objectives must be developed to achieve the overall mission goals.

- A good plan can absorb change in assumptions. The extent to which a plan must be changed if the assumptions are violated is a measure of robustness.

- Actions must achieve stated objectives. A common error is to select targets independent of overall mission objectives. A good plan clearly shows how specific actions support overall mission objectives.

- A good plan has appropriate phasing. The distinction between prioritizing targets and sequencing in an attack is sometimes missed.

Characteristics of good planning done by experienced planners:

- Good planning is done with an appropriate balance between focusing on the mission objectives and an awareness of available resources. Targets should not be on the target list because they could be hit, but rather because they should be hit.

- Assumptions should be explicitly known and kept in the forefront during the planning process. This helps maintain the focus of planning for the assumed context.

- The planning process should have a shorter cycle than the adversary's planning process.

- Good plans are developed at an appropriate level of detail. Without sufficient detail, it is difficult to evaluate whether the plan is feasible. With too much detail, human planners cannot visualize the plan and how it will unfold.

These two sets of characteristics affect how a mixed initiative planning system could be built. Air campaign plans should exhibit the first set of characteristics, whether generated by a human or by an automated or semi-automated planning system. The second set of plan characteristics captures how a good planner develops a plan and may not be necessary for a strictly automated planner. An AI planner may develop plans very differently from how human planners develop plans, taking advantage of the computational power of the computer and of recent advances in AI technology.

However, there may be aspects of good planning that humans excel at and turning the whole task over to an AI planner may degrade plan quality. Thus the importance of grounding the design of a Plan Critic on a Cognitive Task Analysis. If a mixed initiative system is to communicate with the human planner and keep him or her in the loop (in a credible way), then human planning strategies, capabilities, and approaches must be understood. An AI planner will support, not supplant, human planners.

For example, there are at least two important concerns relating to the robustness of a plan. To be robust, a plan has to absorb change. The two primary sources of change are changes in the goals (i.e., commander's intent) and changes resulting from new information acquired during plan implementation, such as discovering that some assumptions were faulty. These two sources of change are qualitatively different and should be handled separately by the Plan Critic.

For a plan to absorb changes in its overall goal, its tasks should accomplish subgoals which are broad enough to cover changes in the overall goal. For example, a plan's primary goal may be the destruction of target A. If the plan accomplishes this goal by first destroying nearby SAM sites, then destroying target A, it will be accomplishing subgoals which support other goals, i.e., the SAM site. Destroying the SAM sites supports the destruction of nearby targets such as targets B & C. If in the middle of plan execution the commander changes the primary target to target B, little effort will have been wasted and only minor modifications to the plan will be necessary.

There are several ways to measure a plan's ability to absorb change in commander's intent. First, the user can specify possible alternative objectives. This does not necessarily need to be an exhaustive list but should be representative of the kind of new objectives expected and should cover the full range of possibilities. The user could be given guidance as to what to consider in specifying alternative objectives. Plans could be generated for these alternative objectives and the degree of overlap with the existing plan calculated. A particular alternative objective may generate several alternative plans. The alternative plan with the maximum degree of overlap with the existing one should be used for the calculation. More simply, the subgoals needed to accomplish the alternative objectives could be extracted from the alternative plans. Robustness could then be based on the percentage of these subgoals accomplished by the existing plan.

To absorb change in the plan assumptions, the overall structure and probability of success of the plan should not rely too heavily on the assumptions. One way of testing this is to change the assumptions and replan. A measure of robustness is then based on the degree to which the plan has changed. The differences between the new plan's and old plan's tasks, their sequencing, and required resources are considered. The replanning can also be performed assuming that some tasks are complete or committed by freezing these tasks during replanning. This is important because the start or completion of some early tasks may commit the force to a particular course of action, making the plan inflexible.

To be considered robust, a plan's probability of success should not diminish greatly when the assumptions change. This computation either requires a simulation to run the plan through or estimates of the probability of successful completion the tasks. The differences in the success probabilities of tasks when assumptions are false is also needed. Given the probabilities, propagating them through the task network to the final objectives is straightforward.

Previous plans in a casebase may also be a useful source of valuable information with which to evaluate a current plan's robustness. An evaluation of a previous similar plan's robustness could be used to help evaluate a current plan's robustness in several ways. For example, a prior plan's evaluation could be used as a basis and adjusted for the current plan. The reasoning behind the evaluation of the similar plan can be applied to the current plan. The process used to evaluate the previous plan can be adapted and used for the current plan.

Similarly, parts of previous plans can be used in the current plan's evaluation. The robustness evaluation for a part of a previous plan which is similar to a part of the current plan can be used as a basis for evaluation and adjustment of the plan. The reasons behind the evaluation of a similar part of a previous plan can be examined to see which might be applicable. The process used to evaluate a similar part of a previous plan might also be applicable. Success probabilities assigned to similar tasks can be used as a basis to estimate the success probabilities for a current task. Alternative objectives for a previous similar plan can be used to suggest possible alternatives to the user for selection. Probabilities assigned to assumptions in previous similar plans can be used as a basis for assumptions in the current plan.

In addition to returning a measure of plan robustness, the Plan Critic should return the rationale supporting its evaluation. This should include both the strengths and weaknesses of the plan. Using this information the user, or the Plan Critic, could then improve the plan. For example, if the Plan Critic reports that a plan is weak because there is one critical subgoal which will only be accomplished if each of a long sequence of events is successful, then the user could develop a second set of events in addition to the first set to accomplish the critical subgoal. Given two independent means of accomplishing the plan's goals, it becomes much more robust.

Conclusion

Our work to date has revealed many characteristics of robust plans and thus requirements for evaluating plan robustness. The next step, apart from continuing to identify characteristics of robustness, is to explore the cognitive processes through which air campaign planners develop and evaluate plans. It is important to consider such processes when building an automated system so that the system is grounded in the user's approach to solving the problem.

There is currently a gap between AI planning technology and the needs of users in the operational community. The goal of this project is to fill that gap by using a design strategy that focuses on the intended user, in this case the air campaign planner, while being vigilant to technological advances. We are using a Decision-Centered Design approach to build a plan evaluation tool, the Plan Critic. This entails conducting a thorough Cognitive Task Analysis of the tasks faced by air campaign planners and incorporating the findings as the foundation of the tool. The result will be a plan evaluation tool that supports the human planning process in an operational setting. Successful development of the Plan Critic will allow air campaign planners to develop robust plans more quickly and effectively.

References

Calderwood, R., Crandall, B. W., and Baynes, T. H. 1988. Protocols of expert/novice command decision making during simulated fire ground incidents, KATR-858-88-02Z. Yellow Springs, OH: Klein Associates, Inc. Prepared under contract MDA903-85-C-0327 for the U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Calderwood, R., Crandall, B. W., and Klein, G. A. 1987. Expert and novice fireground command decisions, AD-B154 224. Alexandria, VA: Defense Logistics Agency.

Crandall, B., and Calderwood, R. 1989. Clinical assessment skills of experienced neonatal intensive care nurses. Yellow Springs, OH: Klein Associates Inc. Final Report prepared for the National Center for Nursing, NIH under Contract No. 1 R43 NR01911 01.

Crandall, B., and Gamblian, V. 1991. Guide to early sepsis assessment in NICU. Fairborn, OH: Klein Associates Inc.

Crandall, B., and Klein, G. A. 1988. Key components of MIS performance. Yellow Springs, OH: Klein Associates Inc. Report prepared for Standard Register, Dayton, OH.

Crandall, B., Klein, G., Militello, L. and Wolf, S. 1994. Tools for applied cognitive task analysis. Final report. The Naval Personnel Research and Development Center, San Diego, CA.

Klein, G. A. 1989. Recognition-primed decisions. In W. B. Rouse (Ed.), *Advances in man-machine system research*, Vol. 5:47-92. Greenwich. CT: JAI Press, Inc.

Klein, G. A., Calderwood, R., and Clinton-Cirocco, A. 1986. Rapid decision making on the fire ground. In Proceedings of the 30th Annual Human Factors Society, 576-580. Dayton, OH: Human Factors Society.

Klein, G.A., Calderwood, R., and MacGregor, D. 1989. Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics* 19:462-472.

Klein, G. A., and Thordsen, M. 1988. Use of progressive deepening in battle management. In Proceedings of the 11th Biennial DoD Psychology Conference, 301-305. Colorado Springs, CO: US Air Force Academy.

Thordsen, M. L., and Calderwood, R. 1989. Protocol analysis of flight crew decision making during simulated malfunctions. Yellow Springs, OH: Klein Associates Inc. Prepared under Order No. A72145C for the NASA/Ames Research Center, Moffett Field, CA.

Thordsen, M., Klein, G. A., Michel, R., and Sullivan, E. 1988. Methods for providing direct feedback about decision processes for command-and-control classroom exercises, KATR-863-8806Z. Yellow Springs, OH: Klein Associates Inc. Prepared under contract MDA903-86-C-0170 for the U. S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Thordsen, M. L., Klein, G. A., and Wolf, S. 1990. Observing team coordination within Army rotary-wing aircraft crews. contract MDA903-87-C-0523 for the U.S. Army Research Yellow Springs, OH: Klein Associates Inc. Prepared under Institute, Aviation Research and Development Activity, Ft. Rucker, AL.

Thordsen, M. L., Militello, L. G., and Klein, G. 1992. Cognitive task analysis of critical team decision making during multiship engagements. Fairborn, OH: Klein Associates Inc. Prepared under PO No. RI-82264X for Williams AFB, AZ.

Warden, J. A. 1988. *The Air Campaign: Planning for Combat*. Washington, DC: National Defense University Press.

Wolf, S., Klein, G., Thordsen, M., & Klinger, D. 1991. Factors affecting the decision making of fire direction officers. Yellow Springs, OH: Klein Associates Inc. Prepared under contract DAAA15-90-C-1054 for the U. S. Army Ballistics Research Laboratory.