Cohen, M.S., Freeman, J.T. and Thompson, B.T. (1998). Critical Thinking Skills in Tactical Decision Making: A Model and A Training Method. (Canon-Bowers, J. and E. Salas, eds.), *Decision-Making Under Stress: Implications for Training & Simulation*. Washington, DC: American Psychological Association Publications.

Critical Thinking Skills in Tactical Decision Making: A Model and A Training Strategy

Marvin S. Cohen, Jared T. Freeman, and Bryan Thompson

Cognitive Technologies, Inc., Arlington, Virginia

Critical Thinking Skills in Tactical Decision Making: A Model and A Training Strategy

Introduction

This chapter describes a model of decision making skills under time stress, a training strategy based on that model, and experimental tests of the training strategy. A prime example of the kind of decision making addressed by the model is Naval ship defense against an approaching aircraft whose intent is ambiguous. The Vincennes and Stark incidents are prototypical cases of this kind, illustrating two different ways that such decision making can go wrong. But similar decision making skills are required in many other domains, such as medicine, fire-fighting, commercial airline emergencies, and even chess. In each case, the decision maker must gauge the available time for collecting and analyzing information and must usually act on an incomplete picture of the situation.

The TADMUS program encompasses several different approaches to this problem. One approach is through workload reduction and automaticity, based on the premise that practice in lower-level skills may free cognitive resources for higher-level decision making. Another approach is through pattern recognition, in which decision makers learn to recognize and respond appropriately to a large number of situations, by-passing deliberate decision making altogether. Still another approach involves training team coordination and communication skills, to ensure that relevant information is communicated to the right people at the right time. These approaches improve decision making by addressing its inputs, the resources it draws on, or alternative processes that make it unnecessary. The approach described here, by contrast, provides a direct focus on decision making processes in novel or unexpected situations.

The second section of this chapter briefly describes a model of real-world decision making under time stress and contrasts it with alternative conceptualizations. The model is based on naturalistic observations of experienced decision makers in real-world contexts and on cognitive theory. The third section of the chapter provides an overview of a training strategy based on the model. The strategy utilizes instruction, demonstration, and practice to teach Naval officers methods for identifying and handling qualitatively different kinds of uncertainty in assessments and plans. The fourth section describes experimental tests of the training at two Naval training facilities with active-duty Naval officers. The effects of training on decision making processes, on situation assessment accuracy, and on appropriateness of action were examined. The final section provides conclusions and lessons learned from the research, including potential extensions of the critical thinking training to other domains.

A Naturalistic Framework for Decision Making

Models of Decision Making

Efforts to train decision making have been shaped by competing conceptions of what decision making is. The approach described here contrasts with, and borrows from, classical decision theory, pattern recognition, and problem-solving approaches to decision making.

Decision Theory. Perhaps the most familiar framework for decision making, classical decision theory, contains two main parts: *Bayesian probability theory* for drawing inferences about the situation, and *multi-attribute utility theory* for selecting an optimal action. Bayesian probability theory requires that decision makers identify a set of mutually exclusive and exhaustive hypotheses (e.g., about the intent of an approaching aircraft). They then assess the probability that each hypothesis is true, identify all the potential observations that might bear on those hypotheses in the future, and quantify the impact each such observation would have. Then, as new observations occur, they can use algorithms from the theory for updating belief in the hypotheses.

Multi-attribute utility theory is an analogous method for choice. It requires that decision makers specify a set of possible actions, an exhaustive and mutually exclusive set of uncertain states of the world, and a set of evaluative dimensions. They then assess the probability of each uncertain state, the importance of each evaluative dimension, and the score of each action-state combination on every evaluative dimension. The theory then enables decision makers to calculate a score reflecting the overall desirability of each action.

We (along with others) have argued that decision theory is not in general *cognitively compatible* with the way experienced decision makers work (Cohen & Freeman, 1996). Problems with the classical framework include the kinds of inputs it demands, the kind of processing it prescribes, and the outputs it produces.

(1) By demanding a complete model up front, with fixed assessments of uncertainty and preference, decision theory tends to discourage the dynamic evolution of problem understanding through time, e.g., as new hypotheses, options, observations, outcomes, and even goals are discovered.

(2) By reducing all uncertainty to a single measure (probability), decision theory obscures important qualitative differences in the way different types of uncertainty are handled, such as gaps, conflict, and unreliable assumptions. Decision theory, for example, treats conflicting evidence the same way it treats congruent evidence, by essentially taking an average. Experienced decision makers, on the other hand, may use conflict as an opportunity to identify the faulty assumptions in their beliefs that produced the conflict (Cohen, 1986). Similarly, decision theory handles conflicting goals the same way it handles congruent goals, by calculating an overall score for each option that is an average of the different goals. Experienced decision makers, by contrast, may try to learn from the conflict, by creating a better option or a deeper

understanding of their true objectives (Levi, 1986).

(3) The output of a decision theoretic model is a statistical average – e.g., 70% chance hostile, 30% chance not hostile – rather than a single coherent picture of the situation. Decision makers cannot visualize, anticipate, or plan effectively for an abstract average.

Decision theoretic models, as they are typically applied, are not only descriptively inadequate, but normatively inadequate as well. As argued in Cohen (1993), appropriate normative principles must capture the relevant qualitative features of the decision making process. If a normative standard is to be used to identify decision making errors, it must be close enough to actual performance for the discrepancies to be interesting.

Pattern recognition. A second approach to decision making skill looks toward more holistic intuition, rather than to formal analysis. It identifies decision making skill with the accumulation, through experience, of a set of virtually automatic responses to recognized patterns. This view has been popular in research on differences between experts and novices, beginning with Chase & Simon's (1973) work on chess. Unfortunately, pattern recognition views say little about the kinds of novel or ambiguous situations that are traditionally associated with the term *decision making*. How is situation assessment accomplished in new and changing circumstances? How are conflicting and unreliable data dealt with? How do decision makers change their minds? When do they stop thinking and act? Although pattern recognition appears to be at the heart of proficient performance, other processes may also often be crucial for success in less routine circumstances. For example, Klein (1993) discusses how options may be tested by mentally simulating their outcomes.

Problem solving. One way to address these difficulties is to define decision making as a special case of problem solving. The decision maker may deploy a range of strategies to find the correct hypothesis or action, for example, dividing the problem into simpler subproblems or working backward from the goal to subgoals and actions that achieve them. Such strategies do not figure prominently in either the decision theoretic or the recognitional approaches.

There are two problems with the traditional problem solving point of view, however. First, it fails to address the central role of uncertainty and risk in decision making (Fischhoff & Johnson, 1990). In this, it resembles recognitional approaches. A second problem is that general-purpose problem-solving does not easily accommodate the role of experience-based recognition. In this, it resembles the decision theoretic approach. The problem-solving approach, as it now stands, is more a promising direction than a full-fledged theory of decision making.

Implications for training. Different conceptions of decision making skill are important because they are associated with different *training strategies*. According to Salas & Cannon-Bowers (1977), a training strategy orchestrates *tools* (such as feedback and simulation) and *methods* (such as instruction, demonstration, and practice) to convey a *content*.

From the decision theoretic point of view, the content of training is a set of general-purpose techniques (Baron & Brown, 1991). The principle tool for defining this content is, of course, decision theory, and the primary method of presentation is explicit classroom instruction. Examples of decision problems are used not as content, but to motivate the formal techniques during instruction, to demonstrate their generality, and for paper and pencil practice (e.g., Adams & Deehrer, 1991).

At the opposite extreme, decision training based on the recognitional point of view attempts to convey examples of decision problems and their solutions, rather than general-purpose techniques. The methods used in recognitional training tend to be demonstration and practice with a large set of illustrative problems rather than explicit instruction, and to incorporate tools like high-fidelity simulation and outcome feedback (Means, Salas, Crandall, & Jacobs, 1993).

The development of an effective training strategy depends on an adequate conception of decision making. If both recognition and uncertainty are important, how are they related? If problem-solving skills are relevant, how do they apply to recognition and uncertainty? The present research has focused on these questions and their implications for training. The training strategy described here builds on recognition, but focuses on problem-solving strategies for dealing effectively with uncertainty.

The Recognition / Metacognition Model

Proficient decision makers are *recognitionally skilled*: that is, they are able to recognize a large number of situations as familiar and to retrieve an appropriate response. Our observations of decision-making performance, in naval anti-air warfare as well as other domains, suggest that recognition is supplemented by processes that verify and improve its results (Cohen, Freeman, & Wolf, 1996). Because of their function, we call these processes *metarecognitional*. Meta-recognition skills probe for flaws in recognized assessments and plans, try to patch up any weaknesses that are found, and evaluate the results.

Metarecognitional skills are analogous to the *meta-comprehension* skills that proficient readers use when they try construct a mental model based on the information in a text. For example, according to Baker (1985), skilled readers continually look for problems, such as inconsistencies or gaps, in the current state of their comprehension, and they adopt a variety of strategies for correcting problems, such as referring back to earlier parts of the text or relating the text to information already known.

To reflect the complementary roles of recognition and metacognition in decision making, we have called this framework the Recognition / Metacognition (R/M) model (Cohen, 1993; Cohen, et al., 1996; Cohen, Freeman, & Thompson, 1997a). Cues in a situation activate an interpretation of their meaning. (For example, an aircraft popping up on radar at high speed, low altitude, and heading toward a U.S. ship from an unfriendly country is recognized as having the intent to attack.) This interpretation in turn may activate knowledge structures that organize actual and potential information. A *story* is one such structure, viz., a causal model that people construct to flesh out assessments of human intent (Pennington & Hastie, 1993). (For example, a story about the attacking airplane may include the motivation of the attacking country, the reason it used this type of airplane, the reason it chose own ship as a target, and what it will do next.) Stories and other structures make up an evolving situation model and plan. According to the R/M model, the integration of observations into such structures may involve a set of metarecognitional processes.

These metarecognitional processes include the following:

1. Identification of evidence-conclusion relationships (or *arguments*) within the evolving situation model and plan. This is simply an implicit or explicit awareness that cue A was *observed* on this occasion, while the assessment (e.g., intent to attack) along with expectations of observing cue B were *inferred*. On some other occasion cue B might be observed and cue A inferred.
2. Processes of *critiquing* that identify problems in the arguments that support a conclusion (e.g., hostile intent) within the situation model or plan. Critiquing can result in the discovery of three kinds of problems: *incompleteness, conflict,* or *unreliability* (Cohen, 1986). An argument is incomplete if it provides support neither for nor against a conclusion of interest. (For example, the kinematics of the track

suggest that it is a military aircraft, but say nothing about the key issue, hostile intent.) Two arguments conflict with one another if they provide support both for and against a conclusion, respectively. (For example, the heading of a track toward own ship suggests hostile intent, while its slow speed argues for routine patrol.) Finally, an argument is unreliable if the support it provides for a conclusion depends on unexamined assumptions. (For example, a track's turning toward own ship suggests hostile intent until it is realized that the track may be too far away to have detected own ship.) Unreliable support may shift or vanish when its premises are further considered.

3. Processes of *correcting* that respond to these problems. Correcting steps may instigate external action, such as collecting additional data, or internal actions that regulate the operation of the recognitional system. These internal actions include attention shifting and assumption revision. Shifting the focus of attention overcomes limitations on the automatic activation of information in memory, and brings additional knowledge into view. Such knowledge may fill gaps in arguments, help resolve conflict, or confirm or deny unreliable assumptions. Adjusting assumptions involves operating the recognitional system as if an uncertain belief were known to be true or false. It permits what-if reasoning, exploratory searching for alternative causes and effects, and eventual adoption of a single coherent model or plan.

4. A higher-level process called the *quick test*, which controls critiquing and correcting. Metarecognitional processing occurs when the expected benefits associated with critical thinking outweigh the costs. It is shaped, like other actions, by past experiences of success and failure. Recognition-based responding will be inhibited

and critical thinking will be initiated when the costs of delay are acceptable, the situation is uncertain or novel, and the costs of an error are high.

Figure 1 summarizes the functional relationships among these processes. A more detailed description of the R/M model may be found in Cohen, Freeman, & Wolf (1996) and Cohen, Parasuraman, Serfaty, & Andes (1997).

A computer-based implementation of the Recognition / Metacognition model has been developed (Thompson, Cohen, & Shastri, 1997). The model, which combines neural and symbolic features, utilizes an adaptive critic architecture for reinforcement learning. The recognitional aspect of the critic learns complex relations that enable it to predict events and trigger appropriate actions. In parallel, the metacognitive critic maintains a model of the recognitional critic and learns appropriate actions for regulating the behavior of the recognitional critic. Machine and human learning experiments in the domain of Naval tactical anti-air warfare are currently being compared in order to test predictions of the R / M model.

Insert Figure 1 about here

Comparison to Other Models

The R / M model contrasts with *classical decision theory* in its approach to the inputs, the processes, and the outputs of decision making. (1) Rather than demanding that all inputs to a model be specified in advance, the R / M model predicts the incremental generation of new hypotheses, options, observations, outcomes, or goals in the course of working the problem. (2) Rather than assigning fixed numerical significance to cues or goals and then mathematically aggregating, R / M permits on-going reinterpretation of cues and goals. (3) Instead of an unrealizable statistical abstraction (e.g., 70% hostile, 30% non-hostile), R / M produces a single concrete picture of a situation, together with a qualitative understanding of its specific strengths and weaknesses.

Metarecognitional processing is a highly dynamic and iterative *problem-solving* strategy. The next processing step is determined locally by the results of earlier steps, rather than by a global, fixed procedure (as in Bayesian inference or in other proposed decision heuristics). Correcting for one problem may sometimes lead to identification and correction of another problem. For example, a gap in a story may be filled by collecting further data or remembering previously known information, or, if these fail, by making assumptions. The resulting more detailed arguments may then turn out to conflict with other arguments. Such conflict may then be addressed by examining the reliability of the conflicting arguments (e.g., shifting attention to the grounds for the arguments). This process stops when the quick test indicates that the benefits of further metarecognitional actions are outweighed by the risks of delay, and that action on the basis of the current best model or plan is called for.

The R / M model reconciles *pattern recognition* with problem-solving. It explains how experienced decision makers are able to exploit their experience-based intuition in a domain and at the same time handle uncertainty and novelty. They construct and manipulate concrete, visualizable models of the situation, not abstract aggregations. Uncertainty is represented explicitly at the metacognitive level, by annotating the situation model or plan to highlight points of incompleteness, conflict, and unreliability. Metacognitive strategies respond to these problems and try to improve the current situation model and plan or find better ones. To quote Dreyfus (1997, p.28), metarecognition is “observation of one’s intuitive practice-based behavior with an eye to challenging and perhaps improving intuition without replacing it…”

Implications for Training

The R/M model yields an approach to training that is distinct from both classical decision theory and pattern recognition models. From the point of view of the R/M model, the content of training is neither a small set of general-purpose methods (like decision theory) nor a vast quantity of specialized patterns and responses. The focus is a moderately sized set of *domain-grounded* strategies for critical thinking. Unlike patterns, these strategies are generalizable, but unlike decision theory, they can only be effectively taught on the basis of pre-existing knowledge in a particular domain.

Several aspects of metarecognitional strategies may be transferable across domains. For example, the same or very similar types of uncertainty (i.e., incompleteness, conflict, and unreliability) seem to be relevant across a variety of domains. Moreover, the same or very similar metarecognitional actions — collecting more data, shifting attentional focus to retrieve more information, and changing assumptions — seem appropriate for handling these types of uncertainty. Finally, the metarecognitional cycle depicted in Figure 1, with its priority of testing first for incompleteness, then conflict, then unreliability, may also be a relatively general tendency among proficient decision makers. Such decision makers try to create a complete and consistent story, and then evaluate the plausibility of the assumptions that it demands. Nevertheless, these generalities presuppose a base of recognitional knowledge upon which they operate. Metarecognitional strategies thus make little sense in abstraction from a particular application area (cf., Kuhn, Amsel and O'Loughlin, 1988).

Training Critical Thinking Skills

This section describes a training strategy for Naval Combat Information Center (CIC) officers that is based on the R / M model. Table 1 outlines the crucial features of this training strategy: its tools, its methods, and its content (Salas & Cannon-Bowers (1977).

Insert Table 1 about here

An essential *tool* in the development of the training strategy is cognitive task analysis. The R / M model and the training design are based on critical incident interviews with active-duty Naval officers, in which they described experiences in the Persian Gulf, the Gulf of Sidra, and elsewhere (Kaempf, Klein, Thordsen, & Wolf, 1996). Our analysis focused on nine incidents in which the officers decided whether to engage a contact whose intent was unknown, under conditions of undeclared hostility. We analyzed the interviews to discover the officers' thinking strategies, ways of organizing information, and decisions. Many aspects of the training are based on differences in the way that more and less experienced officers handled similar types of situations.

The training design utilizes both information-based and practice-based training *methods*. In each segment of training, officers listen to a brief verbal presentation of the concepts central to that segment, followed by questions and discussion. They then play the role of a Tactical Action Officer in realistic scenario-based exercises. Training and test scenarios were adapted from the scenarios described by Johnston (this volume); modifications were made to include events from critical incident interviews that exercise the relevant critical thinking skills.

This scenario-based practice utilizes two important *tools*: interactive simulation and feedback. The simulation platform is the Decision Making Evaluation Facility for Tactical Teams (DEFTT), discussed by Johnston (this volume). Feedback was provided in the form of hints from the instructor in real-time as the scenario unfolded, and by class discussion at the conclusion of the scenario. A third tool consists of training guidelines such as those proposed by Duncan et al.

(1996) for training mental models.

A final tool is represented by a set of performance measures used to evaluate the success of the training. These measures address both critical thinking processes (e.g., the number of pieces of conflicting evidence that trainees identified and the number of factors they considered in arguments about intent), and outcomes (i.e., the agreement of trainees' situation assessments and actions with those of a subject matter expert).

The training *content* is divided into the following four segments:

- *Creating, testing, and evaluating stories.* This section provides an overview of the critical thinking process, called *STEP*. When an assessment is uncertain, decision makers can enhance their understanding by constructing a *S*tory around it. The story includes the past, present, and future events that would be expected if the assessment were true. Decision makers use the story to *T*est the assessment, by comparing expectations to what is known or observed. When evidence appears to conflict with the assessment, they try to patch up the story by explaining the evidence. They then *E*valuate the result. If the patched up story involves too many unreliable assumptions, they generate alternative assessments and stories. In the meantime, they *P*lan against the possibility that their current best story is wrong.
- *Hostile-intent stories*. Stories contain certain typical components. Knowledge of these components can help decision makers notice and fill gaps in the stories they construct. A particularly important kind of story is built around the assessment of hostile intent. For example, a complete hostile intent story explains the motivation for attack, the choice of a target, and the choice of an attacking platform. It also accounts for how that platform localized the target and the manner in which it will arrive at a position

suitable for engaging it. The training teaches officers by practice and example how to discover story components and to let the stories guide them to relevant evidence about intent.

- *Critiquing stories.* After a story is constructed, decision makers step back and evaluate its plausibility. This segment of the training introduces a devil's advocate technique for uncovering hidden assumptions in a story and generating alternative interpretations of the evidence. An infallible crystal ball persistently tells the decision maker that a conclusion is wrong, despite the evidence that appears to support it, and asks for an explanation of that evidence. Regardless of how confident decision makers are in their assessments, this technique can alert them to significant alternatives. It can also help them see how conflicting data could fit into a story. The technique helps decision makers expose and evaluate assumptions underlying their reading of the evidence.
- *When to think more.* Critical thinking is not always appropriate. Officers should not incur extreme risk to own ship rather than engage a track. But it is necessary to evaluate the time available before taking an irreversible action. The decision maker should probably act immediately unless three conditions are satisfied: (1) The risk of delay must be acceptable. (2) The cost of an error if one acts immediately must be high. And (3) the situation must be non-routine or problematic in some way. Training focuses on the way experienced decision makers apply these criteria. For example, they tend to utilize more subtle estimates of how much time is available, based on the specifics of the situation. They focus more on longer-term objectives in estimating the costs of an error. And they show greater sensitivity to the mismatch between the

situation and familiar patterns.

The following four sections describe each of these four training segments in more detail. The discussion, like the training itself, draws on examples based on actual incidents.

Segment #1: Creating, Testing, and Evaluating Stories

Observations regarding a surface or air contact may prompt recognition of its intent. For example, in one incident a Tactical Action Officer (TAO) in the Gulf of Sidra was notified that a track had popped up on radar at close range and was heading toward own ship. Its speed suggested a military plane and it did not respond to radio warning. After progressing two miles, it began to circle. The TAO suspected the aircraft was hostile. However, when the costs of an error (e.g., shooting a friendly) are high and time is available before own ship is at significant risk, it is worth thinking critically about the assessment. This is what the TAO and his Captain did.

Creating a story. Figure 2 outlines four steps of critical thinking, and is called the *STEP* cycle. The first step is to build a *S*tory around the current assessment. Although the term *story* sounds playful, in fact building a story means taking an assessment seriously. The story includes what would have happened in the past, what should be happening now, and what is expected to happen in the future if the assessment is true. The assessment can then be tested by comparing the story to the facts or (if that is not possible) by evaluating its plausibility. The training illustrates the STEP process by examples of increasing complexity drawn from real incidents such as this one.

Insert Figure 2 about here

The situation of the circling aircraft in the Gulf of Sidra did not fit a ready-made attack profile. As a result, the TAO tried to create a hostile intent *Story*, to explain what he had observed and to test the hostile intent assessment. An attack by Libya would fit Ghaddafi's objective of defending his self-proclaimed territorial waters in the Gulf of Sidra. The TAO proceeded to describe what might have happened if the contact was hostile: "I figured that the pilot took off from a Libyan base, kept his head down, and turned directly towards us. He must have wanted to seize the moment...attack just as we detected him and before we got our gear up...."

Testing the story. The next part of the story, however, did not fit what happened: "Instead of continuing to attack, the track paused to circle. This made no sense." Another observation that conflicted with hostile intent was the absence of electronic emissions.

Did the TAO conclude, with a sigh of relief, that the aircraft could not be hostile? Certainly not. Experienced decision makers do not abandon an assessment (especially a threatening one) because it does not fit all the evidence. In many situations, *no* pattern fits all the evidence perfectly — the truth will therefore necessarily run counter to expectations. To give an assessment a fair chance, officers try to incorporate all the observed events into the story, even if at first they don't seem to fit.

In this incident, the TAO tried to fit the conflicting observations into the hostile intent story. "The best interpretation I could make of this -- and it wasn't too good -- was that he came up to target us, but his radar had busted." A single explanation happens to account for both arguments that conflict with hostile intent. The aircraft was not emitting and was not approaching because its radar had broken.

Evaluating the story. Just because a story can be constructed, however, doesn't mean that the story is true. The next step in critical thinking is to *E*valuate: Step back and ask if the story makes sense. It is sometimes possible to gather more data to test an explanation of conflicting evidence. In other cases, it is a matter of rapidly judging plausibility. Did the officer have to stretch believability too much to make all the observations fit? Each time the decision maker explains a piece of conflicting evidence, it is like stretching a spring. Eventually, the spring resists any further efforts in that direction, and snaps back: The “explained away” evidence becomes conflicting again. If time is available and stakes are high, experienced decision makers try to build a different story, based on a different assessment.

In this incident, it was not very plausible that an attacking aircraft would stop and circle in plain view if its radar was not functioning. Since the TAO's story required a stretch, the captain considered the possibility that this was a friendly aircraft.

Another cycle. The captain then generated a new *S*tory based on the assumption that the aircraft was friendly. “The captain... figured it was one of ours, his radio was off or busted, and he was trying to execute our triangular return-to-fort profile [a maneuver to signal a friendly aircraft returning to the battle group].” Unfortunately, when *T*ested, expectations based on this story did not perfectly fit the observations either. The track did not follow the expected triangular profile very closely. The captain did not abandon his assessment, but tried to patch up the story to explain the discrepancy: “That pilot was doing a spectacularly lousy job of drawing that triangle.”

Returning to the *E*valuation stage, we ask, How good is the captain’s story? Like the TAO’s hostile intent story, it requires the assumption of broken equipment (radar or radio, respectively). In addition, it assumes a poorly executed maneuver. This, however, seemed more plausible than the TAO’s assumption that a hostile aircraft with a broken radar would stop to circle. The captain accepted the assessment that the aircraft was probably friendly. As the TAO noted, “The captain was right.”

Planning against weaknesses in the story. Although the captain believed the aircraft was friendly, he knew that he might be wrong. Therefore, he *P*lanned against this possibility by continuing to monitor the aircraft's behavior and readying relevant weapons systems.

The second and third segments of the training delve into specific aspects of the STEP cycle. In particular, the second segment looks at a particular kind of story, for hostile intent. The third segment, on critiquing, discusses methods for helping fit discrepant observations into a story for generating alternative assessments, and for making plans more robust.

Segment #2: A Hostile Intent Story Template

Stories based on the assessment of hostile intent take on special importance when own ship is being approached by a contact whose purpose is unclear. In these situations, there is a consistent set of issues that experienced decision makers tend to consider. Figure 3 provides a causal structure, or template, for a hostile intent story that incorporates those issues.

The central element in this structure is the current intent of the enemy: to attack with a particular asset (or assets) against a particular target (or targets). The left side of the structure represents prior causes of the intent, and the right side represents the effects or consequences of the intent in the current situation. The point of telling a story is not simply to fill the slots. It is to try to *make sense of*, or *argue for*, the hostile intent assessment from the vantage point of each of these causes and effects.

Insert Figure 3 about here

The causal factors at the left in Figure 3 (goals, opportunity, and capability) make up what might be called "the big picture," and they often shade the way kinematic cues are interpreted. Interviews suggest that more experienced officers try to create a complete story about hostile intent, incorporating all these factors. Less experienced officers tend to be more myopic: They often focus only on past and present kinematics, i.e., on the speed, altitude, range, and heading of a track, rather than on the larger context and future predictions.

The training employs examples from the interviews and simulated scenarios for practice with each component of the hostile-intent story template.

Higher-level goals. In what way is the country owning the platform motivated to attack a U.S ship? Evidence that a country's *goal* is to attack U.S. ships can take different forms, which vary in strength: from prior incidents of actually engaging U.S. ships, to intelligence regarding an incipient attack, to public threats and increased tensions. In one practice scenario, for example, trainees usually decide not to engage rapidly approaching F-4s from Iran, based on their understanding of the rules of engagement (ROE). Later in the same scenario, they are fired upon by an Iranian boat. Still later they are again rapidly approached by a similar group of F-4s. This time, most choose to engage the F-4s, even though the literal application of ROE to this situation is the same as before (when they chose not to shoot). When such overt evidence is lacking, officers often rely on prior intelligence regarding a planned attack. In several incidents, officers cited the lack of such prior intelligence as a key factor in causing them to doubt the hostile intent of a contact.

Opportunity. How is own ship a logical target for attack given the country's goals and other potential targets that it could have chosen? Opportunity can be a subtle cue regarding intent. In one practice scenario an air contact on its way toward own ship passes a U.S. command and control ship. The latter is at least as lucrative a target as own ship and is more accessible for attack. The failure to select the most favorable target does not disprove hostile intent, but argues against it. The failure may have a good explanation — for example, the contact did not detect or have intelligence regarding the command and control ship, or there is some unsuspected reason that own ship is a more desirable target. A complete hostile intent story must include some such explanation, which must then be tested or judged for plausibility. Conversely, the presence of a lucrative target such as a flagship provides support for, but does not prove, the assessment of hostile intent.

Capability. How is the track a logical choice as an attack platform given the goals and available assets of the attacking country? In a number of incidents officers puzzled over the employment of less capable platforms, such as a gunboat, helicopter, or light aircraft, against a U.S. AEGIS cruiser. Again, this argues against hostile intent but does not disprove it: The conflict may have a plausible explanation, such as unwillingness to sacrifice expensive resources or willingness to conduct a kamikaze raid.

Localization. How did the attacking platform detect own ship, or how is it attempting to do so? Localization is a surprisingly frequent concern among experienced officers, and heavily influences the reading of more standard kinematic cues. For example, officers tend to regard a contact that emerges from a hostile nation, turns toward own ship, and speeds up as hostile. But if this contact is too far away to have detected own ship, these cues must mean something else. The hostile intent story may be patched up, perhaps by assuming third-party targeting support or improved equipment or training. These explanations can then be tested or evaluated. Conversely, in other incidents, when a track appeared too slow or too high to fit a hostile profile, its behavior was sometimes reconciled with a hostile intent story by assuming a need to localize a target.

Generalizing the story concept. Stories contain characteristic events associated with an intent. This idea can be generalized to assessments other than hostile intent. In a form of discovery learning (Collins & Stevens, 1983), we ask trainees to imagine that one of the aircraft in a practice scenario is on a search and rescue mission. We then ask them to tell a story around that assessment, which includes past, present, and future events. As we record the events volunteered by trainees on a whiteboard, and draw causal arrows between them, a set of typical components and relationships emerges. Stories typically specify goals, opportunities to achieve goals, methods or capabilities for achieving the goals, actions preparatory to achieving the goal, and execution of the intended action.

For example, the components of the search and rescue story produced by one class included the following: (1) Opportunity: There is something in the water to be rescued. (2) Goals: Rescue is not overriden by risks such as on-going combat. (3) Capability: The organization and capabilities are present to mount a search and rescue operation in the time available (i.e., the rescue is not so fast as to seem staged). The most appropriate available platform for search and rescue is chosen. (4) Arriving in position: The platform's speed, altitude, and flight pattern are appropriate for search. The contact will indicate by radio response that its mission is search and rescue. (5) Execution: The contact is engaging in, or will engage in, actions appropriate for rescue.

A similar story can be built around any assessment of intent, for example, a lost friendly aircraft, harassment, provocation, or attack, each of which has its own set of characteristic components.

As the STEP process implies, stories are not fixed patterns or checklists associated with particular intents. Stories are constructed when no pattern fits all the observations. A story template is filled in differently each time it is used, depending on how the decision maker tries to explain the conflict. The uniqueness of stories makes the evaluation step crucial. No single element is conclusive by itself: The decision maker must look at *the whole story*. Hostile intent is supported when the available information fits easily within the hostile intent template, or when the assumptions required to make it fit are tested and confirmed. The information weighs against hostile intent if a large number of unusual and untested assumptions are needed to make it fit.

Segment #3: Critiquing Stories

Overcoming overconfidence. Once a story has been constructed, the decision maker evaluates its plausibility. The simplest approach to evaluation is direct: Just ask, for example, "How confident am I that this platform intends to attack my ship?" Decision theory requires that answers be given in the form of numerical probabilities. There is evidence, however, that this approach can be seriously misleading. People, including experts, tend to be overconfident when they provide direct probabilistic estimates of confidence (Lichtenstein, Fischhoff, & Phillips, 1982). Overconfidence, in turn, can cut thinking short before key issues have been explored, and it may be one reason for unfortunate surprises or overhasty decisions.

A more useful approach to the estimation of confidence does not presuppose that the decision maker is already aware of all the factors contributing to uncertainty. In the R / M model, confidence is assessed by actively ferreting out the important assumptions underlying an assessment. In the remainder of this section, we will describe a powerful strategy for identifying assumptions, called the Crystal Ball Technique. Then we will describe the role it can play in various phases of the STEP cycle.

The Crystal Ball Technique. An argument may appear plausible at first blush, but have weaknesses that are not immediately recognized. Such weaknesses may be revealed by deliberate critiquing. A powerful strategy for exposing weaknesses is for decision makers to *imagine* that the arguments for a conclusion are false — i.e., that the observations are true but the conclusion is wrong — and search for an explanation. Each explanation generated in this way reflects an assumption underlying the original argument; that argument is valid only if no such explanation is true.

This segment of the training introduces a devil's advocate method, called the Crystal Ball Technique, that consists of four steps:

1. Select an important conclusion, such as the intent of a contact, or the belief that a track cannot localize own ship at its present distance.
2. Imagine that an infallible crystal ball (or, equivalently, a perfect intelligence source) tells you that this conclusion is wrong — despite the observations (or reports or analyses, etc.) that suggested it was true.
3. Explain how this could happen, i.e., how the conclusion could be wrong despite the evidence supporting it. How does this change the way the evidence is interpreted?
4. Optionally, the crystal ball now tells you that this explanation is wrong and sends you back to step 3, to devise another possible explanation of the evidence. Continue until the set of exceptions to your original conclusion seems representative of the ways the conclusion could be wrong.

Testing an assessment. In the Test phase of STEP, the decision maker identifies arguments *against* an assessment, and then uses the Crystal Ball Technique to critically probe these arguments. As a result, decision makers identify assumptions that would produce a coherent story despite apparently conflicting observations.

The assessment may be the decision maker's favored hypothesis, or it might be an alternative hypothesis that happens to be under consideration. In both cases, the Crystal Ball forces decision makers to explain how the assessment can be true despite conflicting evidence. In the former case, the Crystal Ball can prevent decision makers from giving up a favored hypothesis prematurely. In the latter case, the Crystal Ball forces the decision maker to seriously consider how an alternative assessment might account for evidence that appeared to support the favored hypothesis. In both cases, the crystal ball helps the decision maker go beyond the initial, recognized interpretation of evidence, to find possible alternative meanings. In both cases, the assessment can then be evaluated by looking at the assumptions required to save it.

Consider, as an example, an incident in which an officer tried to patch up a hostile intent story. An aircraft was approaching an AEGIS cruiser on a straight course from the direction of Iran at slow speed and low altitude. The aircraft was not emitting or responding to IFF challenges. Every one of these observations – flying toward U.S. ships, emerging from Iran, low altitude, failure to respond to IFF – suggested an attack, except slow speed.

The Anti-Air Warfare Officer (AAWC) had a possible explanation for some of these observations. Friendlies might appear to originate from Iran and fail to respond to IFF challenges if they had turned off their IFF transponders as a precaution before flying near hostile territory and then forgot to turn them back on when coming out. (In some cases deliberate critiquing of an argument is not necessary. Conditions under which the argument fails may be so familiar that they are recognized virtually at the same time as the argument itself.)

The AAWC, however, continued to take seriously the possibility that the aircraft was hostile. To create a coherent hostile intent story, it was necessary to find an explanation of its slow speed. The officer forced himself to imagine that the aircraft was hostile despite its slow speed. This led to the consideration and rejection of a series of possible explanations:

Candidate assumption #1. Slow speed could be consistent with hostile intent if the aircraft were trying to locate its target. This explanation could be tested. If the aircraft were trying to locate a target visually, it would be flying a search pattern rather than a straight course. But it wasn't ("He's not scanning visually for anybody because that's a straight line").

Candidate assumption #2. Since this explanation was disconfirmed, the AAWC forced himself to generate another one. The aircraft might be flying toward own ship if it had prior intelligence on shore regarding the location of the target ("I wonder what their intelligence capability is?... Do they know where I am?"). However, this prediction was also at least partly disconfirmed, since the aircraft was not flying directly toward own ship.

Candidate assumption #3. The AAWC then forced himself to find still another explanation: Perhaps the aircraft was trying to locate a target electronically. But in that case, its radar would be emitting detectable signals— which it wasn't.

Candidate assumption #4. The only explanation the AAWC could think of now was that the aircraft planned to shoot blind! ("I can't imagine him shooting blind. They could, though — they did it once, shot something off and hoped it hit something... So you're thinking, well, he probably won't shoot, but he might.") This explanation was a last-ditch effort to save the assumption that the aircraft is hostile. It could not be directly tested, but was judged implausible.

Because he had failed to construct a plausible hostile intent story, despite heroic efforts, the AAWC delayed engagement of the approaching contact. Just in case it was hostile, however, he warned it with fire control radar. The aircraft immediately turned on its IFF transponder and squawked a friendly response.

Evaluating an assessment. In the Evaluate phase of STEP, decision makers step back to ask if a story makes sense. In doing so, they judge the plausibility of the assumptions required by the story. In this phase, they may also generate alternative assessments and begin a new STEP cycle. In subsequent Evaluation phases, they can compare the assumptions they must accept if the different assessments are true.

The Crystal Ball Technique can be used to probe the arguments *for* a favored assessment. The crystal ball says that a hypothesis other than the favored one can account for the evidence, and forces decision makers to generate such hypotheses. Koriat, Lichtenstein, & Fischhoff (1980) found that overconfidence was reduced by forcing subjects to search for reasons why they might be wrong.

We employ an actual incident as a classroom exercise, in which students use the crystal ball both to flesh out a hostile intent story and to discover possible non-hostile intents. An AEGIS cruiser, escorting a flagship through the Straits of Hormuz, detected two Iranian fighters taking off from an Iranian air base and identified them as F-4s from a brief radar transmission. Typically, these aircraft would be expected to head north or south along the Iranian coast for a routine patrol. In this case, however, the planes began circling the airport. These circles gradually widened until the aircraft were coming within their weapons range of the cruiser. While circling, the aircraft kept their search radar on continuously. As the circles widened, the aircraft switched to fire control radar, locking on the cruiser during the portion of the orbit when they were pointed toward it. At the point in each orbit when the lock was broken, the pilot of one of the aircraft switched back to search mode.

The observation of circling aircraft in broad daylight leads to the recognitional conclusion that their intent is not hostile. In a normal attack, according to the captain, the F-4s would “come screaming at me” fast and low. But here they were, “in broad daylight; they know we’re here, we know they’re there.”

Nevertheless, this behavior did not fit the pattern of a routine patrol. The Captain therefore tried to put together a hostile intent story. In fact, a disturbing number of elements for such a story were present. There was appropriate motivation — Iranian hostility to the U.S. —

and appropriate capability — F-4s armed with anti-ship missiles. There was also appropriate opportunity: The presence of the flagship “obviously heightened our interest,” according to the captain. Localization could be explained by the deliberate use of search radar. (Turning on this radar was noteworthy, the captain said, because, “The Iranians did not have the maintenance capability to [routinely] fly their electronics and burn them steadily.”) Gradually widening circles and locking on put the aircraft in position to engage. The captain noted that by illuminating own ship, the aircraft had already satisfied the rules of engagement (ROE).

As a class exercise, the instructor tells the trainees that an infallible crystal ball has determined that the aircraft are hostile, despite the observation that they are circling in broad daylight; and the trainees are asked to explain how this could be. This typically elicits a number of potential explanations of the circling: for example, the aircraft might be planning to fire a standoff weapon. The crystal ball now says that the aircraft have hostile intent, but they do not intend to fire a standoff weapon. Other suggestions are now forthcoming: for example, the aircraft may intend to divert the cruiser from attack by other aircraft or surface vessels. This process, repeated several times, elicits still more suggestions. For example, the aircraft may be waiting to rendezvous with other aircraft for a concerted attack. The aircraft may be waiting in order to synchronize their activity with other aircraft. They may be targeting for other aircraft. Their radar may work better at high altitude. They may be having rudder or communications problems; and so on.

Knowledge of these possibilities provides the captain a real basis for evaluating the plausibility of the hostile intent assessment. If the captain believes the aircraft to be hostile, he must be prepared to assume that at least one of the explanations (or some similar one) is the case.

Next, the crystal ball declares that the aircraft are not hostile, and asks for explanations of the circling pattern and the emissions. A variety of explanations are forthcoming here as well: For example, the aircraft might be circling as part of a test of the radar; their radio may be broken and they may be signaling the airport; they may be testing U.S. Rules of Engagement; they may be harassing U.S. ships; and so on. If the captain believes the aircraft are not hostile, he must be willing to accept one of these explanations, or some similar one. In this incident, the captain concluded that the intent of the aircraft was to harass rather than to attack his cruiser.

Planning against weaknesses in an assessment. The crystal ball exercise, though it takes only a few minutes, can have very practical consequences. Many of the explanations can be tested. In the example of the circling F-4's, explanations implying diversion or coordination with other aircraft or ships in the area may lead to heightened vigilance, and may be either confirmed or disconfirmed by observations. Other explanations (e.g., those regarding the aircrafts' weapons or radar characteristics) may be confirmed or disconfirmed by intelligence. Some explanations may be dismissed as implausible. Some can provide the basis for contingency planning in case they turn out to be true. In this incident, the captain developed a contingency plan for the possibility of attack, by ordering internal ship defensive systems to a high state of readiness.

Generalization of critiquing skills. The most important result of the crystal ball exercise is a heightened appreciation of the fallibility of judgment. Trainees are typically surprised at the number of ways an apparently compelling argument can fail (e.g., the argument that circling aircraft do not intend to attack). After about 30 minutes of practice with different scenarios, trainees start considering exceptions to conclusions spontaneously, without explicitly invoking the device of a crystal ball.

The crystal ball should reduce overconfidence, but its purpose is not to *undermine* confidence. Rather, it teaches decision makers that *all* assessments require assumptions of some sort. The existence of alternative interpretations, or of conflicting evidence, is not proof that an assessment is wrong. By encouraging decision makers to compare the assumptions (or stories) implied by different assessments, it helps them select the best overall account. After the crystal ball exercise, officers may believe their original conclusion even more strongly than before — or they may change their minds. But the best way to *earn* confidence in an assessment is to take seriously the possibility that it is wrong.

The crystal ball method can be used to uncover hidden assumptions in plans, as well as stories. In this case, the crystal ball attacks the connection between actions and goals instead of the connection between evidence and conclusions. It says that the planned actions will be attempted but the goals of the operation will not be achieved, and demands an explanation. This can lead to a greater understanding of the weaknesses in the current plan. The result may be a changed plan, an elaborated set of contingencies, acceptance of risk, or greater confidence in the existing plan.

Segment #4: When to Think More

Critical thinking is not always appropriate. Yet decision makers in combat cannot afford valuable time thinking about whether to think. What we call the *quick test* is used to decide rapidly and without excessive overhead when to critique and improve an assessment and when to go ahead and act on it. The quick test requires a balance among the costs of delay, the costs of error if one acts without further critical thinking, and the degree to which the situation is either unfamiliar or problematic (Cohen, Parasuraman, Serfaty, & Andes, 1997). Experienced officers seem to differ from less experienced officers in the way they approach each of these judgments.

Costs of delay. Less experienced officers typically base judgments of available time on the enemy's doctrinal weapon's range. As soon as own ship is within range of the contact's weapons these officers stop thinking and are ready to shoot. By contrast, more experienced officers do not settle for stereotypical or doctrinal estimates of weapons range. They buy more time for decision making by considering factors that are specific to this enemy and to this situation. For example, they consider history (at what ranges have they in fact launched in past training or combat, rather than how far does the manual *say* they can shoot?), or visibility conditions on that occasion (is it a moonless night?). They may also consider actions the enemy must take prior to launching (such as changing altitude or communicating), which will tip them off that an attack is imminent. At the same time, they may develop tripwires and contingency plans to make own ship's response to an attack as rapid as possible.

In sum, experienced officers explicitly ask themselves, "How much time do I have before I must act?" And they buy time for decision making by estimating available time more precisely, and planning more carefully, than less experienced officers.

<u>Stakes of an error.</u> Less experienced officers tend to focus on immediate goals, such as survival of own ship and destruction of hostiles. More experienced officers give more consideration to higher level and longer-term stakes. For example, they place more weight on avoidance of an international incident or other organizational objectives, including damage to their own career.

<u>Situation typicality.</u> An unidentified aircraft heading toward own ship at high speed (i.e., the expected kinematics of an attack) gets the attention of a junior officer just as quickly as the attention of a senior officer. Less experienced officers also learn patterns such as commercial air schedules, corridors, and speeds, or routine patrol routes, speeds, and altitudes, and they are very good at detecting behavior that fits or does not fit such patterns.

More experienced officers, by contrast, are more sensitive to the possibility that a situation may not fit *any* pattern perfectly. They are less likely to conclude that a contact intends to attack simply because it fails to match a routine patrol checklist and does match a few aspects of an attack pattern (e.g., heading and speed). They are quicker to notice that it *also* fails to match the attack pattern perfectly (e.g., the platform is unable to localize own ship from that distance; this is not a logical target for that platform). In sum, more experienced decision makers realize that a situation is ambiguous in cases where less experienced decision makers do not.

Experimental Tests

Critical thinking training has now been tested at two Navy training facilities. Results of the two studies will be described together. However, some important differences between them are summarized in Table 2. Study 1 was conducted at the Surface Warfare Officers School (SWOS), Newport, RI, while study 2 was conducted at the Naval Post Graduate School (NPS), Monterey, CA. The two studies represent a tradeoff between the number and experience level of participants (Study 1 was better) versus the duration and quality of training (Study 2 was better).

Insert Table 2 about here

In both studies, training focused on decision making skills of individual officers, although they received the training in classes of five or six. We are now testing a team version of the training at SWOS.

Method

Design

Both Study 1 and Study 2 involved a pretest-treatment-posttest design, and both studies used two scenarios, which were counterbalanced between pretest and posttest across groups. In addition, Study 1 varied the treatment condition (training vs. control) across groups. Study 2 did not utilize a control group.

Participants

Sixty officers at SWOS participated in Study 1 (40 in the training condition, 20 as controls). All were in the Navy, with an average of 9.5 years of military service. These officers were being trained to serve as department heads in engineering, operations, or weapons. Ninety-two percent had performed shipboard duty in the Combat Information Center (CIC), and 32% had served as Tactical Action Officer (TAO).

At NPS, 35 officers took part in Study 2. As at SWOS, these officers averaged 9.5 years of military service. However, only 51% were Navy, while 14% were Marines, 29% were Army, and 6% were Air Force. Forty-six percent of the officers at NPS had worked in the CIC or in similar tactical positions in the Marines, Air Force, or Army. Only 14% of officers at NPS had served specifically as TAO.

Materials

The critical thinking training materials used in these experiments included a training text, brief explanatory lectures, discussions, and exercises. In study 1, all practice scenarios were presented for class discussion by the instructor. In study 2, four practice scenarios were simulated on the Decision Making Evaluation Facility for Tactical Teams (DEFTT) (see chapter by Johnston, this volume). These scenarios were modified to make critical thinking more appropriate, i.e., to reduce the total number of tracks while increasing uncertainty about the intent of some of the remaining tracks. In most cases, these modifications replicated the situations that had been described in critical incident interviews. Participants performed in these scenarios individually, acting as TAOs. Feedback from the instructor was provided in real-time as the scenario unfolded and, after it concluded, in group discussion.

The pretest and posttest scenarios in both studies were DEFTT simulations. These simulations began with oral and written briefings concerning the geopolitical context of the scenario and the military forces involved. Each participant then turned to a personal computer that simulated a command and display (C&D) station. The C&D presents symbology concerning the identity, speed, bearing and range of air and surface tracks, as well as textual details concerning these characteristics and the track's response to electronic interrogation (IFF). Virtually all tracks in these scenarios except own ship and accompanying surface craft were symbolically marked as unknown (rather than as friend or foe) and were unresponsive to electronic interrogation. Audio communications were presented to simulate internal and external comms. Most communications concerned the location of tracks, their presumed identity (e.g., F-4, Mirage), and electronic warfare (EW) data received from the tracks (such as search radar or fire control radar emissions). In study 2 these communications were edited so that they could be understood by non-Navy officers. In study 1 EW data were also provided on a large display in the center of the classroom.

Participants performed scenarios in groups of five or six, but each participant worked independently. Participants could not consult with their classmates during tests nor take any actions that would alter events in the scenario for themselves and their classmates. Specifically, they could not maneuver own ship, fire weapons, interrogate tracks, or initiate communications during scenarios, though they could indicate their intent to perform such actions in response to questions during test breaks.

Procedure

For participants in study 1, the experiment began with discussion and practice to re-familiarize them with the DEFTT. For participants in study 2, who were generally less experienced, the experiment began with a presentation concerning the function of the Combat Information Center (CIC), the role of the Tactical Action Officer (TAO), and the operation of the DEFTT simulator.

The pretest and posttest scenarios were each paused at three points. During each break, participants received a test booklet consisting of five questions:

(1) Assess the intent of a single [experimenter-designated] track and defend that assessment.

(2) Generate alternative possible intent assessments and estimate confidence in each.

(3) Select an assessment of the designated track that you do not agree with and defend it.

(4) Identify evidence that conflicted with an [experimenter-designated[intent assessment and then defend the assessment.

(5) Describe actions the participant would take at this time in the scenario.

Participants did not know which track would be the focus of attention or which intent assessment would be designated for consideration, until after the relevant segment of the scenario was completed and the break began.

These test questions specifically described the desired behavior. For example, participants were asked to identify and explain conflicting evidence, and to generate alternative assessments, rather than simply to evaluate a hypothesis. The intent was to mitigate any effects of experimenter demand on the results. The expected behavior should be as clear on the pretest (and in the control group) as in the post-training test. Any differences in performance should thus be attributable to the effects of training on the relevant meta-recognitional skill.

Measures based on these questions covered every phase of the STEP process: Stories — the number and variety of issues considered when evaluating an assessment in questions 1 and 3;

Test — the amount of conflicting evidence that a participant identified and the number of explanations generated to patch up stories in question 4; Evaluate — the number of alternative assessments generated in question 2 and the accuracy of the assessment favored by the participant in question 1; and Plan — the frequency with which contingency plans were created in question 5.

Following the pretest, participants received training, except for members of the study 1 control group. The latter completed a psychological battery, listened to a lecture on problem solving and knowledge representation strategies, and discussed challenging problems in their jobs as weapons officers, engineers, or operations officers. They appeared to find the control condition enjoyable and interesting. After the training or control treatment, participants executed the posttest. In addition, all participants completed a biographical survey concerning military experience, and all except controls evaluated the training.

As noted above, the experiments differed in the duration and realism of training. In study 1, pretest, training, and posttest occurred in a single day, with only 90 minutes for training itself. By the time of the posttest, signs of fatigue were evident. In study 2, on the other hand, events were broken into five two-hour sessions over two weeks. (The introduction to the CIC and DEFTT occupied the first session, the pretest occurred in the next session, two training sessions followed, and the posttest was administered in the final session.) Training in study 2 utilized DEFTT scenarios, and was therefore both more realistic and more similar to test conditions than study 1.

Results

Successful critical thinking training should have an impact both on decision processes and, through such processes, on the accuracy of decisions. Table 3 summarizes the main results from both studies. In study 1, where training was shorter and the posttest came at the end of a long day, we found either trends or significant effects on all of the critical thinking skills. In study 2, training had a significant effect on all the critical thinking skills. A more complete description of the results for study 1 may be found in Cohen, Freeman, Wolf, & Militello (1995).

Insert Table 3 about here

Effects of Training on Decision Processes

Filling out stories. One of the objectives of training with story templates is to increase the scope of the factors that officers consider when evaluating an assessment. The factors considered should go beyond present and past kinematics of the track to also include goals, opportunity, capability, localization, deceptive aspects of the approach (i.e., alternative interpretations of kinematics), and predictions regarding future kinematics. This analysis pertains to questions 1 and 3, in which officers were asked to defend an assessment they favored and an assessment they did not favor, respectively.

In study 1, there was a trend for training to increase the number of factors considered in an assessment of intent. The number of arguments generated by trained participants was 6.5% greater for favored hypotheses (question 1) and 8.6% greater for disfavored hypotheses (question 3), compared to untrained participants (for questions 1 and 3 combined, $F_{1,47} = 2.953$, $p = .092$).

These effects became much larger in the more favorable training and testing regime of study 2. Training increased the number of arguments participants presented in defense of their favored assessments (question 1) by 22%. The number of arguments grew from a mean of 5.12 per break on the pretest to 6.26 on the posttest ($t_{33} = 3.807$, $p = .001$). The percentage effect was larger when trained participants defended assessments they did not agree with (question 3): an increase of 43% from a mean of 3.01 per break on the pretest to 4.31 on the posttest (t_{33} = 3.807, p = .001).

Did training simply increase the quantity of issues considered, or did it also influence the type and variety of issues that officers thought about? We analyzed the distribution of arguments across story components in study 2. Training reduced the percentage of arguments that reflected present and past kinematics from 61% on the pretest to 51% on the posttest, and correspondingly increased the percentage of arguments reflecting other factors (χ_1^2 = 10.816, p = .001). Taking these other factors separately, we found an increase due to training in every non-kinematics story component: i.e., goals, capabilities, opportunity, localization, deceptive features of the track's approach, and predicted future actions.

As the quantity of arguments increased, did quality decline? If so, training might simply lower the threshold for reporting or thinking about an issue, rather than expanding the scope of understanding. This, however, was not the case. A subject matter expert (the retired naval officer who designed the test scenarios) did a blind rating of the relevance and impact of each argument provided by each participant in study 2. There was no effect of training on the average quality of arguments. The mean quality of arguments was 5.3 on the pretest and 5.4 on the posttest (on a 10-point scale).

Identifying conflicting evidence. Another important objective of training was to improve an officer's ability to identify evidence that conflicts with an assessment. Question 4 specifically asked subjects to list evidence that conflicted with an assessment designated by the experimenter.

In study 1, participants who received training identified 52% more items of conflicting evidence than controls. Trained subjects identified an average of 1.4 items per break while controls identified 0.9 items ($F_{1,55}$ = 6.236, p = .015). Training increased the amount of conflicting evidence identified whether or not these participants happened to agree with the experimenter-designated assessment (Agreement was determined by comparing the designated assessment with the assessment favored by the participant in question 1). In study 2, training boosted the amount of conflicting evidence identified by 58%, from an average of 1.6 items on the pretest to 2.6 on the posttest ($t_{32} = 5.481$, $p < .001$).

Explaining conflicting evidence. In ambiguous and complex situations, almost any assessment conflicts with some of the evidence. Yet one assessment, however implausible it may seem, must turn out to be true. To discard an assessment simply because there is evidence that conflicts with it, then, would mean rejecting all assessments and never finding the truth. More constructively, conflict can be taken as a cue to think more deeply about assumptions underlying one's interpretation of the evidence. Conflicting evidence may point to an exceptional circumstance that explains the conflict. Question 4 not only asked officers to identify evidence that conflicted with an experimenter-designated assessment, but also to defend the assessment by generating explanations of the conflicting cues.

In study 1 trained officers generated 70% more explanations (.679 explanations per posttest break) than controls (.400 explanations per posttest break). However, variation between test scenarios was quite large, and so this positive pattern was not statistically reliable. In study 2, however, training boosted the number of explanations significantly, by 27%, from 2.566 on the pretest to 3.250 on the posttest ($t_{32} = 4.920$, $p < .001$).

Generating alternative assessments. Generating alternative assessments is an important part of evaluating a story. By suggesting alternative interpretations of observations, it exposes hidden assumptions in the current story, which can be tested or judged for plausibility. In some cases, an alternative assessment may be found which is better than the current hypothesis.

In study 1, there was a trend for trained participants to generate more alternative assessments than controls. Officers who received training generated 9% more assessments per break (3.6, on average) than controls (3.3) ($t_{59} = 1.498$, $p = .140$).

The effect of the more extensive training in study 2 was highly reliable. The number of alternative assessments generated on a given break rose 41%, from 2.689 on the pretest to 3.792 after training ($t_{34} = 5.880$, $p < .001$). This increase in quantity was not accompanied by a decrease in quality. The subject matter expert's blind rating of the plausibility of assessments fell a non-significant 3% between pretest and posttest ($t_{34} = 0.567$, $p = .574$).

Contingency planning. The final phase of the STEP process is to plan against the possibility that one's favored assessment is wrong. Such planning is a way of buying time for critical thinking or for collecting more data.

In study 1 actual engagements were rare among both trained and control participants. However, trained participants were twice as likely as controls to identify explicit contingencies or tripwires for engagement. An average of 6% of the control participants developed contingency plans for engagement on each break, but 13% of the trained participants did so ($F_{1,57} = 8.362$; $p = .005$). (Planning was not analyzed for study 2.)

Confidence in assessments. The training successfully teaches officers to question assumptions, notice conflicting evidence, and generate alternative assessments. It is natural to worry that such training would diminish officer's confidence in their assessments of enemy intent and their decisiveness in taking action. However, this is only a surface view of what the training is designed to accomplish. First the officers are taught that critical thinking is appropriate only under special circumstances, where time, stakes, and uncertainty warrant it. Second, they can stop at any time and act on their current best assessment. Third, and on a deeper level, the training gives officers a better understanding of the reasons for confidence in an assessment. Trainees are taught that even though no story is perfect, some story, however imperfect, will turn out to be true. Hence, training emphasizes the importance of evaluating and selecting among stories, and it shows trainees how this can be done. Exploring the assumptions underlying assessments should lead to the conclusion that the assessment ultimately chosen, while imperfect, is the best available.

As a metric of confidence, we took the difference between confidence ratings for the two assessments in which a participant expressed the most confidence on Question 2. This reflected the subject's ability to discriminate between the preferred assessment and the second best. In study 1, confidence ratings rose 12.5% with training. While not a statistically reliable increase, this indicates at the least that training did not *lower* confidence. Moreover, in study 2 confidence ratings rose 20% from pretest to posttest, a marginally significant result ($t_{33} = 1.985$, $p = .055$).

Effects of training on decision quality

The findings above demonstrate that training based on the R/M model alters the ability of officers to generate, defend, and rebut assessments. However, it does not speak to the ultimate outcome: Do these critical thinking processes increase the accuracy of situation understanding and enhance the success of actions guided by that understanding?

As a first step in this analysis, we examined whether training changed the types of assessments officers generated. The assessment in which each subject was most confident was categorized as either *hostile*, *not hostile*, or *unknown*. For officers in both studies, training reliably affected the category of assessment that subjects preferred, broken down by scenario and test break (study 1, $\chi^2_8 = 24.17$, $p = .002$; study 2, $\chi^2_6 = 24.05$, $p = .001$).

We evaluated the quality of assessments by (1) comparing them with the assessments of a subject matter expert (SME) and (2) by measuring consensus among subjects. In addition, we examined whether the actions officers proposed changed with their assessments.

Accuracy of assessments. The standard of accuracy was the assessment of tracks at each break by the retired senior Navy officer who designed the scenarios. In both experiments, training produced large improvements in accuracy on one of the two test scenarios, but no change in either direction in the other. We focus on the scenario in each experiment that elicited effects. 77% of the trained officers in study 1 were in agreement with the assessments of the SME, compared with only 43% of controls, an improvement of 79% ($\chi^2_2 = 6.337$, $p = .013$). Among officers in study 2, training boosted agreement with the SME by 35% — from 60% on the pretest to 81% on the posttest (although the increase was significant at break 1 only: $\chi^2_2 = 6.791$, $p = .034$).

Consensus. An alternative index of accuracy is the level of consensus among subjects regarding their assessments. Training that improves accuracy should raise consensus among subjects as they converge on a common interpretation of events. We used as a measure of consensus a metric from information theory, called "average uncertainty," (Garner, 1962) which is defined as:

$$U = - \Sigma p(x) \log (p(x))$$

Here, $p(x)$ is the relative frequency with which members of the group picked hypothesis x. U is zero when members of a group all agree, and grows larger with disagreement.

Training appeared to increase consensus among trained officers in both studies. Among officers in study 1, average uncertainty was 14% lower overall with training ($U = 0.911$) than without (1.042). Training lowered average uncertainty 41% among officers in study 2, from $U = 0.31$ to 0.22.

Actions. Assessments of the intent of a track may be expected to influence actions. In the Study 1 scenario that elicited training effects on assessment quality, the intent of the approaching track could have been participation in a search and rescue (SAR), or the SAR operation may have been a cover to close on own ship and attack. Controls were more likely than trained officers to assess tracks as hostile, and they took actions that reflected this, such as vectoring CAP and illuminating the tracks ($F_{1,28} = 2.635$, $p = .081$). Trained participants (and the SME) were more likely to offer assistance in the search and rescue ($F_{1,28} = 3.382$, $p = .077$). In sum, training improved the accuracy of situation assessments, and officers' actions changed accordingly. (Actions were not analyzed in study 2.)

Subjective evaluations of training

The participants in both experiments provided quantitative and qualitative evaluations of the training. During debriefing, subjects rated the training on a scale from 1 (strongly negative) to 5 (strongly positive).

Seventy-three percent (29 of 40) of the officers in the trained group of study 1 gave the instruction a positive rating (4 or 5). There were seven neutral ratings (3), four negatives (2), and no strongly negative ratings (1). The average rating among participants in study 1 was 3.7. Officers with tactically oriented specialties (weapons and operations) gave the training a higher rating than engineers or deck officers ($F_{1,38} = 4.055$, $p = 0.051$).

In study 2, 71% (25 of 35) of the participants rated the training positively (4 or 5). Six participants were neutral (3), two were negative (2), and one was strongly negative (1). The average rating by officers in study 2 was 3.7. Officers with some tactical experience in the Navy or other military services tended to rate the training more positively than officers with no such experience.

Qualitative evaluations of the training were also similar for the two studies. Most participants found the training useful in solving the test problems and anticipated that it would be useful in the field. For example, participants said the training would help "organize what I have been doing previously and take it to another level," "stop me from making assumptions," "reinforce the concept that the obvious answer may not be the correct answer," and "keep tunnel vision to a minimum." Participants mentioned favorably the processes of organizing information in stories and using the devil's advocate to generate alternative interpretations of evidence.

Lessons Learned: Guidelines for Implementing Critical Thinking Training

Lesson Learned 1: Development of critical thinking training starts with cognitive task analysis.

While the overall strategy of critical thinking training is generalizable across domains (see Lesson Learned 7 below), the training must be based on substantive content that is specific to a particular domain. Analysis of critical incident interviews is a valuable tool for identifying critical thinking requirements in a domain, in particular, by comparing the performance of decision makers at different experience levels. The same critical incidents can be used for generating demonstration, practice, and test materials for training in that domain.

Lesson Learned 2: Critical thinking skills are important in tactical decision making.

Critical incident interviews provide evidence for the importance of critical thinking skills in proficient tactical decision making. More experienced decision makers differed from less experienced ones along a variety of dimensions. These skills include: going beyond pattern matching in order to create plausible stories for novel situations, noticing conflicts between observations and a conclusion, elaborating a story to explain a conflicting cue rather than simply disregarding or discounting it, sensitivity to implausible assumptions in explaining away too much conflicting data, ability to generate alternative stories, planning against the possibility that the current assessment is wrong, and more careful attention to the time available for decision making. These critical thinking skills presumably help experienced decision makers handle uncertainty effectively without abandoning the recognitional abilities they have built up.

Lesson Learned 3: Effective critical thinking training combines instruction with realistic practice.

An effective strategy for training these critical thinking skills combines information-based instruction on critical thinking concepts, demonstration of critical thinking processes, and guided practice in realistic problems. Instruction sensitizes trainees to the concepts to be learned and helps them assimilate lessons during practice. Simulation exercises provide an opportunity to demonstrate critical thinking skills and to provide feedback in real time. Such experience is crucial if trainees are to learn to use critical thinking skills in a realistic, time-stressed situation.

This strategy is effective in teaching critical thinking processes. In the second of two studies, training increased the number of factors officers considered in assessing the intent of a track by 30%, increased the amount of conflicting evidence they noticed by 58%, increased the number of assumptions they identified underlying that evidence by 27%, and increased the number of alternative assessments they generated by 41%.

There is evidence that this critical thinking training strategy can improve not only processes but outcomes. Agreement of assessments with a subject matter expert increased significantly in two out of four test scenarios in the two studies, by 79% and 35%, respectively. At the same time, the training tended to increase officers' confidence in their assessments. In addition, most subjective evaluations of the training were positive.

Lesson Learned 4: Critical thinking training may be taught in segments.

Both critical thinking instruction and practice can be conveniently divided into segments. The first segment provides an overview of critical thinking processes, and introduces the idea of building, testing, and evaluating a story, and planning against its weaknesses. The second segment focuses on particular kinds of story (such as the hostile intent story). The third segment provides a devil's advocate strategy for evaluating a story and finding alternative interpretations of observations. A fourth segment provides strategies for deciding when to think critically and when to act immediately.

<u>Lesson Learned 5. Critical thinking training is reasonably robust over training conditions.</u>

Some benefits of critical thinking training were observed even with a very short period of time for training, and with testing and training compressed into a single day (Study 1). However, greater benefits were realized when there was more time for training, and the training and testing were spread out (Study 2).

<u>Lesson Learned 6: Critical thinking training can be extended to teams.</u>

A team-oriented training strategy for critical thinking is now being tested at SWOS. The training content includes two specifically team-based skills: (1) Team leaders are taught to articulate periodic situation updates (Entin, Serfaty, & Deckert, 1994) that mention problems with assessments, such as missing, unreliable, or conflicting evidence. Such updates not only provide team members with a *shared mental model* of the tactical situation (Entin, Serfaty, & Deckert, 1994), but also foster a *shared metacognitive model* of on-going uncertainties in the situation model. The shared metacognitive model prompts team members to volunteer relevant information or insights. (2) Team members are taught to play the part of devil's advocates (crystal ball) for one another, to generate new assessments and interpretations of evidence.

<u>Lesson Learned 7: The critical thinking training strategy is generalizable across domains.</u>

A similar training strategy has been applied in the development of critical thinking training for Army battlefield command staff (Cohen, Freeman, & Thompson, 1997a). This domain differs in a variety of ways from Naval anti-air warfare. In particular, significantly more time is available for decision making, and far more complex plans can be generated.

Critical thinking training has also been applied to the interaction of users with decision aids, in the Rotorcraft Pilot's Associate program. This training focuses on techniques for monitoring the reliability of decision aid conclusions, and the use of different interaction strategies as a function of trust in the aid, stakes, and available time (Cohen, Parasuraman, Serfaty, & Andes, 1997). Recent research explored applications of the model to the design of decision aids for automated target recognition (Cohen, Thompson, & Freeman, 1997). Another domain where critical thinking training is currently being explored is commercial airline pilot decision making.

Successful application in these domains suggests that the critical thinking training strategy is transferable across domains. Future research should explore potential applications to business and other non-military applications more intensely, further test the R / M model, apply it to decision aid design, and explore potential synergies between critical thinking training and decision support.

ACKNOWLEDGMENTS

We wish to thank Al Koster, Robin Waters, and John Poirier of Sonalysts, Inc.; Bill Kemple of the Naval Post Graduate School; and Steve Wolf and Laura Militello of Klein Associates for their help during this research.

References

Adams, M. J. & Feehrer, C. E. (1991). Thinking and Decision Making. In J. Baron & R. V. Brown (Eds.), Teaching Decision Making to Adolescents (pp. 79-94). Hillsdale, NJ: Erlbaum.

Baker, L. (1985). How do we know when we don't understand? Standards for evaluating text comprehension. In D. L. Forrest-Pressley, G. E. MacKinnon, & T. G. Waller (Eds.), Metacognition, cognition, and human performance (Vol. 1) (pp. 155-205). NY: Academic Press.

Baron, J. & Brown, R. V. (Eds.). (1991). Teaching Decision Making to Adolescents. Hillsdale, NJ: Erlbaum.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), Visual information processing (pp. 215-281). NY: Academic Press.

Cohen, M. S. (1986). An expert system framework for non-monotonic reasoning about probabilistic assumptions. In J. F. Lemmer & L. N. Kanal (Eds.), Uncertainty in artificial intelligence (pp. 279-293). Amsterdam: North-Holland Publishing Co.

Cohen, M. S. (1993). The naturalistic basis of decision biases. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), Decision making in action: Models and methods (pp. 51-99). Norwood, NJ: Ablex.

Cohen, M. S. & Freeman, J. T. (1996). Thinking naturally about uncertainty. In Proceedings of the Human Factors & Ergonomics Society, 40th Annual Meeting. Santa Monica, CA: HF&ES.

Cohen M. S., Freeman, J. T., & Thompson, B. T. (1997a). Training the naturalistic decision maker. In C. E. Zsambok & G. Klein (Eds.), Naturalistic decision making, (pp. 257-268). Mahwah, NJ: Erlbaum.

Cohen M. S., Freeman, J. T., & Thompson, B. T. (1997b). Integrated critical thinking training and decision support for tactical anti-air warfare. Third International Command and Control Research and Technology Symposium Proceedings.

Cohen, M. S., Freeman, J. T., & Wolf, S. (1996). Meta-recognition in time stressed decision making: Recognizing, critiquing, and correcting. Human Factors, 38 (2), pp. 206-219.

Cohen, M. S., Freeman, J., Wolf, S., & Mitelino, L. (1995). Training metacognitive skills in naval combat decision making. Arlington, VA: Cognitive Technologies, Inc.

Cohen M. S., Parasuraman, R., Serfaty, D., & Andes, R.. (1997). Trust in Decision Aids: A Model and a Training Strategy. Arlington, VA: Cognitive Technologies, Inc.

Cohen, M. S., Thompson, B., Freeman, J.T. (1997). Cognitive aspects of automated target recognition interface design: An experimental analysis. Arlington, VA: Cognitive Technologies, Inc.

Dreyfus, H. L. (1997). Intuitive, deliberative, and calculative models of expert performance. In C. E. Zsambok & G. Klein, (Eds.), Naturalistic decision making (pp. 17-28). Mahwah, NJ: Erlbaum.

Duncan, P. C., Rouse, W. B., Johnston, J. H., Cannon-Bowers, J. A., Salas, E., & Burns, J. J. (1996). Training teams working in complex systems: A mental model-based approach. Human Technology Interaction in Complex Systems, 8, pp. 173-231.

Entin, E. E., Serfaty, D., & Deckert, J. C. (1994). Team adaptation and coordination training (Tech. Rep. No. 648-1). Burlington, MA: Alphatech, Inc.

Garner, W. R. (1962). Uncertainty and structure as psychological concepts. New York: Wiley.

Kaempf, G. L., Klein, G., Thordsen, M. L., & Wolf, S. (1996). Decision making in complex command-and-control environments. Human Factors, 38 (2).

Klein, G. A. (1993). A Recognition-Primed Decision (RPD) model of rapid decision making. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), Decision making in action: Models and methods (pp. 138-147). Norwood, NJ: Ablex.

Koriat, A, Lichtenstein, S., & Fischhoff, B. (1960). Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 6, pp. 107-118.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). The development of scientific thinking skills. San Diego: Academic Press, Inc.

Levi, Isaac (1986). Hard Choices. Cambridge: Cambridge University Press.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In Kahneman, D., Slovic, P., & Tversky, A. (Eds.), Judgment under uncertainty: Heuristics and biases (pp. 306-334). Cambridge: Cambridge University Press.

Pennington, N., & Hastie, R. (1993). A theory of explanation-based decision making. In G.A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), Decision making in action: Models and methods (pp. 188-201). Norwood, NJ: Ablex.

Salas, E. & Cannon-Bowers, J. A. (1997). Methods, tools, and strategies for team training. In M. A. Quinones and A. Ehrenstein (Eds.), Training for a rapidly changing workplace: Applications of psychological research (pp. 249-279). Washington, D. C.: APA Press.

Thompson, B.T., Cohen, M.S., & Shastri, Lokendra (1997). A Hybrid Architecture for Metacognitive Learning (Hybrid Architectures for Complex Learning Program, Office of Naval Research Contract No. N00014-95-C-0182). Arlington, VA: Cognitive Technologies.

Table 1

Methods, Tools, and Content of the Critical Thinking Training Strategy.

Strategy	Tools	Methods	Content
Critical thinking training (for individuals)	• Cognitive task analysis (critical incident interviews) • Simulation (DEFTT, with specifically tailored scenarios) • Feedback (from group & instructor) • Training guidelines • Performance measures (process & outcome)	• Information-based: lecture and discussion • Practice-based: Guided practice, behavior modeling	• Building stories in novel situations • Detecting and handling conflicting evidence • Generating and evaluating alternative assessments • Adjusting to the available time

Table 2

Differences between Study 1 and Study 2.

Feature	Study 1	Study 2
Location	Surface Warfare Officers School, Newport, RI	Naval Post Graduate School, Monterey, CA
Number of participants	n = 60	n = 35
Experience of participants	All active-duty Navy	Half in Navy, half in other services
Design	Trained (40) versus Control (20) x pretest-posttest	Pretest-posttest (no control group)
Duration of training	90 minutes	4 hours
Practice tools	Paper and pencil (though testing involved simulation)	Simulation (DEFTT)
Scheduling	All training & testing in a single 8-hour session	Training & testing in 2-hour sessions over 5 days

Table 3

Summary of the Effects of Training in Studies 1 and 2 (Averaged Across Test Scenarios).

Variable	**Study 1**	**Study 2**
Number of issues considered re assessment	+ 7%	+ 30%
Number of conflicting pieces of evidence identified	+ 52%	+ 58%
Number of explanations of conflict generated	+ 26%	+ 27%
Number of alternative assessments generated	+ 10%	+ 41%
Accuracy of assessment (agreement with SME)	+ 42%	+ 18%
Consensus on assessment	+ 14%	+ 41%
Confidence in assessment	+ 13%	+ 20%
Frequency of contingency planning	+ 217%	[not analyzed]
Subjective evaluations of training	73% positive	71% positive

Figure Captions

Figure 1. Components of the Recognition / Metacognition model.

Note. Adapted with permissions from Human Factors, Vol. 38, No. 2, 1996. Copyright 1996 by the Human Factors and Ergonomics Society. All rights reserved.

Figure 2. The STEP process for critical thinking.

Figure 3. Hostile intent story template.

Note. Reprinted with permissions from Human Factors, Vol. 38, No. 2, 1996. Copyright 1996 by the Human Factors and Ergonomics Society. All rights reserved.

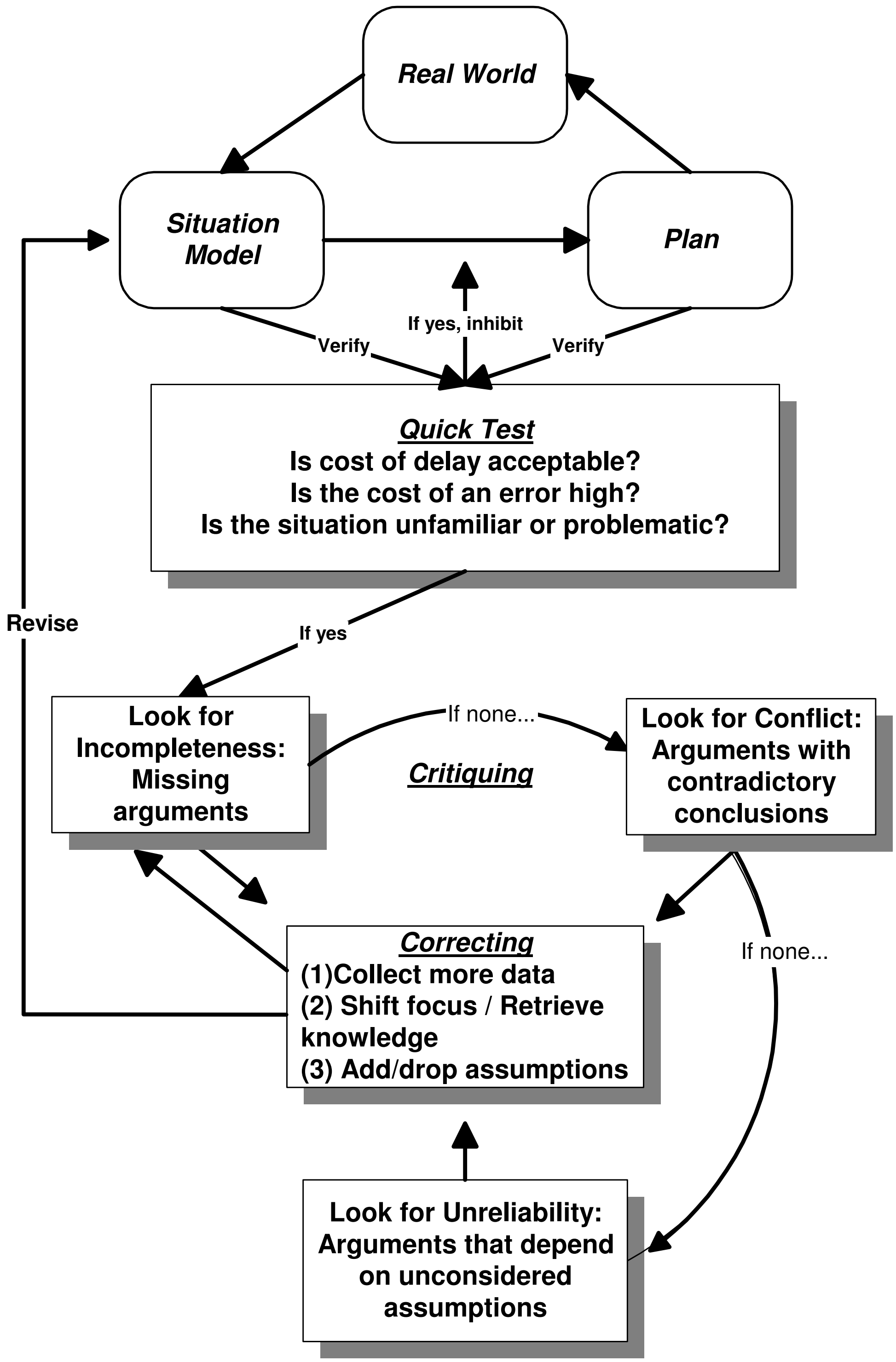
Real World
Situation Model
Plan
If yes, inhibit
Verify
Verify
Quick Test
Is cost of delay acceptable?
Is the cost of an error high?
Is the situation unfamiliar or problematic?
Revise
If yes
Look for Incompleteness: Missing arguments
If none...
Critiquing
Look for Conflict: Arguments with contradictory conclusions
If none...
Correcting
(1)Collect more data
(2) Shift focus / Retrieve knowledge
(3) Add/drop assumptions
Look for Unreliability: Arguments that depend on unconsidered assumptions

Figure 1

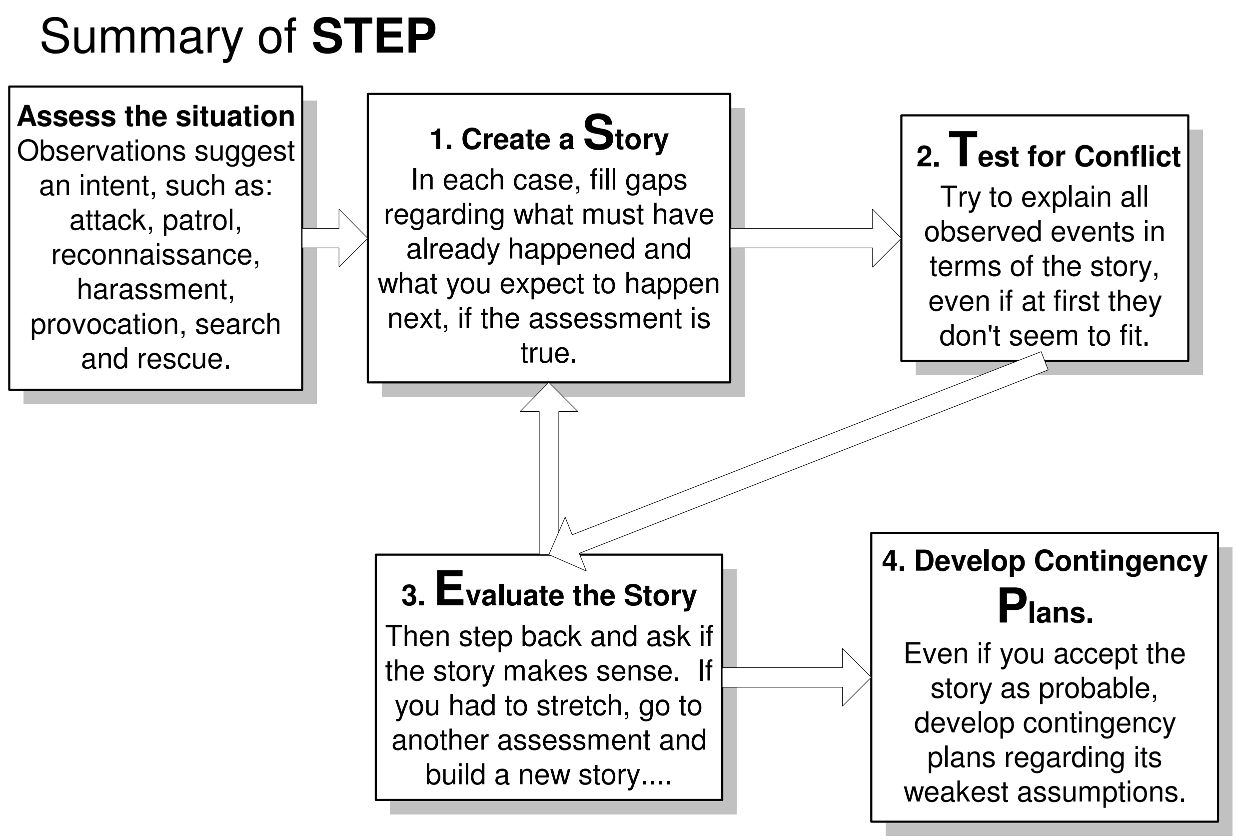
Summary of STEP
Assess the situation
Observations suggest an intent, such as: attack, patrol, reconnaissance, harassment, provocation, search and rescue.
1. Create a Story
In each case, fill gaps regarding what must have already happened and what you expect to happen next, if the assessment is true.
2. Test for Conflict
Try to explain all observed events in terms of the story, even if at first they don't seem to fit.
3. Evaluate the Story
Then step back and ask if the story makes sense. If you had to stretch, go to another assessment and build a new story....
4. Develop Contingency Plans.
Even if you accept the story as probable, develop contingency plans regarding its weakest assumptions.

Figure 2

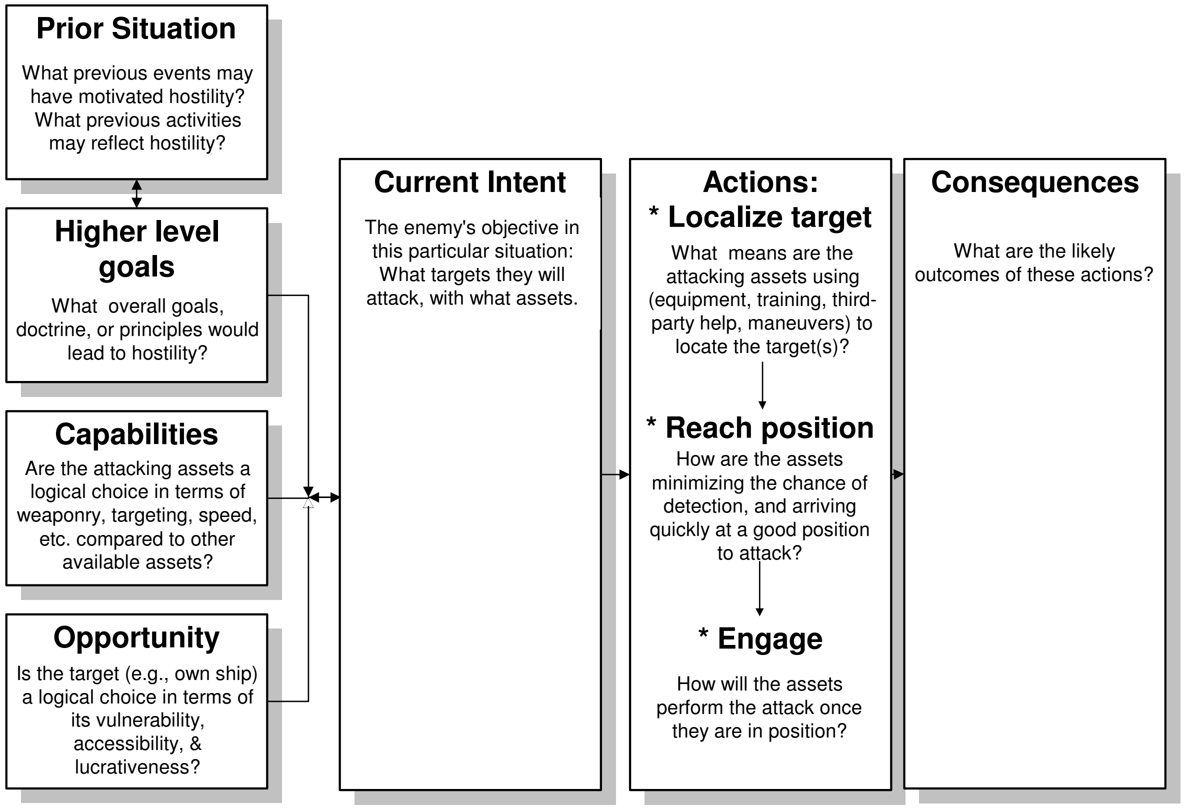
Prior Situation
What previous events may have motivated hostility? What previous activities may reflect hostility?
Higher level goals
What overall goals, doctrine, or principles would lead to hostility?
Capabilities
Are the attacking assets a logical choice in terms of weaponry, targeting, speed, etc. compared to other available assets?
Opportunity
Is the target (e.g., own ship) a logical choice in terms of its vulnerability, accessibility, & lucrativeness?
Current Intent
The enemy's objective in this particular situation: What targets they will attack, with what assets.
Actions:
* Localize target
What means are the attacking assets using (equipment, training, third-party help, maneuvers) to locate the target(s)?
* Reach position
How are the assets minimizing the chance of detection, and arriving quickly at a good position to attack?
* Engage
How will the assets perform the attack once they are in position?
Consequences
What are the likely outcomes of these actions?

Figure 3