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

Efforts to train decision making have been shaped by competing conceptions of what decision making is. The most familiar approach utilizes Bayesian decision theory. It includes methods for creating mathematically consistent judgments and decisions, by exhaustively identifying hypotheses, evidence, and action outcomes, and quantifying their relationships. We (and many others) have argued that decision theory is not in general *cognitively compatible* with the way experienced decision makers work (Cohen & Freeman, 1996). By demanding complete models up front, with fixed assessments of uncertainty and preference, decision theoretic models discourage the dynamic evolution of problem understanding through time. By reducing all uncertainty to a single measure, probability, they obscure important qualitative differences between different types of uncertainty (such as gaps, conflict, and unreliable assumptions). When different sources of evidence appear to point toward conflicting hypotheses, for example, decision theoretic models essentially take an average. By contrast, experienced decision makers use the conflict as an opportunity to reexamine assumptions— for example, about the reliability of the conflicting sources (Cohen, 1986). Finally, the output of a decision theoretic model is a statistical average, which cannot be visualized or planned for, rather than a single coherent picture of the situation. Because of this cognitive incompatibility, decision theoretic models may also be inappropriate as *normative* standards for the evaluation of decision making performance (Cohen, 1993). A more appropriate evaluation of decision making performance would be based on normative principles that capture the relevant qualitative features of the decision making process.

A second approach looks in a different direction for the nature of decision making skill, toward the accumulation with 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 decision making in novel or ambiguous problems. 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 recognition appears to be at the heart of proficient decision making, other processes may also often be crucial for success. For example, Klein (1993) discusses how options may be tested by mentally simulating their outcomes.

Each conception of decision making skill is associated with a different training *strategy*. 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 for structuring and quantifying evidence, hypotheses, options, and outcomes (Baron & Brown, 1991). The principle tool for defining this content is, of course, decision theory, and the primary method of presentation is explicit instruction. Examples as content play a largely secondary role, to motivate the formal techniques during instruction, to demonstrate their generality, and for practice with paper and pencil problems (e.g., Adams & Deehrer, 1991). At the opposite extreme, decision training based on the recognitional point of view attempts to convey a large set of patterns characteristic of actual expert practice in a particular domain. The methods used in recognitional training tend to be demonstration and practice with a large set of examples rather than explicit instruction, and to incorporate tools like high-fidelity simulation and outcome feedback (Means, Salas, Crandall, & Jacobs, 1993).

Other contenders have entered the fray. It is tempting, for example, to try 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 are largely omitted from both the decision theoretic and 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 essential role of experience-based recognition. In this, it resembles the decision theoretic approach.

If both recognition and uncertainty are important in decision making, 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 effective training. We have developed a strategy for training that builds on recognitional processes, is qualitatively compatible with the performance of proficient decision makers, and yet deals 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 *meta-recognitional*. Meta-recognition skills probe for flaws in recognized assessments and plans, try to patch up any weaknesses that are found, and evaluate the results. They 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 test the current state of their comprehension, and they adopt a variety of strategies for correcting problems that are found, such as inconsistencies or gaps in their understanding.

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, 1997). In the R / M model, the basic level of cognition is recognitional, including processes that activate assessments in response to internal and external cues. For example, an aircraft popping up on radar at high speed, low altitude, and heading toward a U.S. ship from an unfriendly country will probably suggest an intent to attack. Assessments in turn may be associated with knowledge structures that organize actual and potential information into a situation model or plan. A *story* is a special type of causal model that people construct to understand human intent (Pennington & Hastie, 1993). For example, an assessment of intent to attack suggests events that must have happened in the past — including

motives, beliefs, and preparatory actions like detecting and localizing a target. This assessment will also suggest events that are expected in the future, like continuing to approach and launching a missile, and actions that may be taken in response. These events and their causal connections constitute a *story* based on hostile intent. According to the R/M model, the integration of observations into situation models and plans often occurs under the influence of meta-recognitional processes.

Meta-recognitional processes include:

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 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 does not provide support either for or against a conclusion of interest. (For example, the kinematics of the track suggest only that it is a military aircraft, but say nothing about hostile intent; this conclusion is too general for deciding whether to engage.) 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 it provides support for a conclusion, yet the support depends on unexamined assumptions. 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, and two kinds of internal actions, attention shifting and assumption revision, that regulate the operation of the recognitional system. Shifting the focus of attention stimulates retrieval of new, potentially relevant information in

long-term memory and brings additional arguments into view for meta-recognitional critiquing. Adding or dropping assumptions permits what-if reasoning, queries for alternative causes and effects, and adoption a single coherent model or plan. These processes in combination help to fill gaps in models or plans, resolve conflict among arguments, and search for arguments that are more reliable.

4. A higher-level process called the *quick test*, which controls critiquing and correcting. Metarecognitional strategies, like other actions, are shaped in part by past experiences of success and failure. Metarecognitional processing occurs when the benefits associated with critical thinking outweigh the costs. This is likely to be the case when the costs of delay are acceptable, the situation is uncertain or novel, and the costs of an error in acting on the current recognitional conclusion are high. In other words, time is available for critical thinking, there is room for improvement in recognitional conclusions, and the stakes are worth it. The quick test considers these three factors and, if conditions are appropriate, inhibits recognition-based responding and interposes a process of critical thinking. When these conditions are not satisfied, the quick test allows immediate action based on the current best response.

Figure 1 summarizes the relationships among these processes. It highlights the functional distinction between recognitional processes (at the top of the figure) and metacognitive ones (the shaded boxes). The recognitional level *provides information* to the metacognitive level, while the metacognitive level *exerts control* over the recognitional level. In the R/M model, metacognition monitors recognitional processing in order to maintain a model or description of it; this model includes an identification of arguments and problems of incompleteness, conflict, and unreliability. When problems are found, metacognition modifies recognitional activity by inhibiting overt action, by shifting attention, and by adopting or dropping assumptions. These functional differences between recognition and metacognition may or may not correspond to structural or physiological ones (see Nelson & Narens, 1994). A more detailed description of

the R/M model may be found in Cohen, Freeman, & Wolf (1996).

Insert Figure 1 about here

The R / M model explains how experienced decision makers are able to exploit their experience 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 (such as 70% chance of hostile intent, 30% chance non-hostile). Uncertainty is represented explicitly at the metacognitive level, by "annotating" the situation model or plan to highlight points of incompleteness, conflict, and unreliability. In response to specific problems of this kind, metacognitive strategies try to improve the current situation model and plan or find better ones.

Metarecognitional processing is highly dynamic and iterative. The next processing step is determined locally by the results of earlier steps, rather than by a global, fixed procedure (as in Bayesian inference). Correcting for one problem may sometimes lead to identification and correction of another problem. For example, a gap in an argument may be filled by collecting further data or remembering previously known information, or, if these fail, by making assumptions. The resulting more specific argument may then turn out to conflict with other arguments. Such conflict may then be addressed by looking for unreliability in one of the conflicting arguments. In doing so, metarecognitional processing might shift focus from the conclusion to the grounds of the argument. This may result in retrieval of previous experiences with the source of the information that is the grounds for the conflicting argument. Such experiences may suggest that the source is not to be trusted. The conflict, which arose because of the implicit, or unconsidered, assumption that this source was accurate, is now resolved. (Alternatively, what if no relevant information were retrieved about the source? A new cycle of critiquing would identify this gap in knowledge, and it might be corrected, for example, by adopting the tentative assumption that the unfamiliar source is not trustworthy. Conflict would be eliminated, but the story now depends on the

potentially unreliable assumption about this source. Attention might now be shifted to the other conflicting 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 output is a coherent, consistent model or plan together with an understanding of its strengths and weaknesses.

From the point of view of the R/M model, the content of training is neither a small set of general-purpose methods nor a vast quantity of specialized patterns and responses. The focus is a moderately sized set of *domain-grounded* strategies for critical thinking. Several aspects of meta-recognitional 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. Moreover, the meta-recognitional 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 general features leave much room for domain-specific variation. Such variation can occur in the types of arguments that are examined first, in the pattern and sequence of corrective actions, in the types of assumptions that are considered acceptable, and in the standards for judging reliability. Thus, meta-recognitional strategies make little sense in abstraction from a particular application area (cf., Kuhn, Amsel and O'Loughlin, 1988). Domain-specific metarecognitional learning builds on the recognitional skills of each domain, and is inextricably tied to those skills. We suspect that meta-recognitional skills can only be trained effectively when situated in the context of a particular application area.

Dreyfus (1977, p. 28) summarizes this idea very well in his concept of *deliberative rationality*: "…whereas most expert performance is ongoing and nonreflective, the best of experts, when time permits,

think before they act…. Deliberative rationality is detached, reasoned observation of one's intuitive practice-based behavior with an eye to challenging and perhaps improving intuition without replacing it…."

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 *content* is divided into four segments: (1) An overview of the cycle of creating, testing, and evaluating stories to improve situation understanding. (2) A particular kind of story based on hostile intent. (3) Strategies for handling conflicting evidence and for generating alternative interpretations of evidence. (4) Guidelines for deciding when critical thinking is appropriate and when immediate action is necessary. To convey this content, the training design utilizes both information-based and practice-based training *methods*. In each segment, officers listen to a brief verbal presentation of the concepts central to that segment, followed by questions and discussion. They then participate in realistic scenario-based exercises designed to provide practice in the relevant skill. This practice utilizes two important *tools*:

interactive simulation, and feedback provided by group discussion and by the instructor. Training guidelines, such as those proposed by Duncan et al. (1996) for training mental models, also may be regarded as tools in the design of this training. A final tool is represented by a set of performance measures that we used to evaluate the success of the training, both in teaching critical thinking processes and in improving outcomes.

We call the training strategy *critical thinking* because it is designed for situations where familiar patterns or rules do not fit. For example, some features of the situation may match the standard hostile intent pattern (e.g., an aircraft turning toward own ship from a hostile country). Other features, however, may not match the standard pattern (e.g., its speed is slower than expected), and may even match parts of another pattern (e.g., commercial airliner). In such situations, more experienced officers explicitly ask themselves how much time they have before they must commit to action. How long until the risk of waiting becomes unacceptable? In many situations, when the contact is an *immediate* threat, the critical thinking strategies are not appropriate. (Nothing in this training implies that officers should incur extreme risk to own ship rather than engage a track.) However, it is also important not to undertake an irreversible action, such as engaging a potential threat, before it is necessary to do so. The bulk of the training focuses on how the available time can be most effectively spent.

The four training segments are the following:

- *Creating, testing, and evaluating stories.* This section provides an overview of the critical thinking process, called *STEP*. When an assessment is uncertain, decision makers take it seriously 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 begin the cycle again. 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 why an attack is taking place against a particular target by a particular platform. It also accounts for how that platform has localized the target and is arriving 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 the current assessment 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 successfully alert them to significant alternatives. It can also help them see how conflicting data could fit into a story. In each case, 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. Unless three conditions are satisfied, the decision maker should probably act immediately: (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 precise estimates of how much time is available, based on the specifics of the situation. They adopt a longer-term

outlook 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 sketch each of these four training segments in more detail. All examples are based on actual incidents elicited from naval officers (Kaempf, Klein, Thordsen, & Wolf, 1996).

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. Figure 2 outlines four steps of critical thinking, and is called, appropriately, 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 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 *T*est 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....” 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 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 dangerous one) because it does not fit all the evidence. In many situations, *no* pattern fits all the evidence perfectly — the truth will 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.

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 a quick judgment of plausibility. Did the officer have to stretch believability too much to make all the observations fit? If so, and if time is available and stakes are high, experienced decision makers try to build a different story, based on a different assessment. 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 evidence that had been explained resumes its original interpretation as conflicting. 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. The captain then generated a story 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, 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.” 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.

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.”

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

A Hostile Intent Story Template

Stories based on the assessment of hostile intent occupy a place of 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 return to, over and over. 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. As shown in Figure 3, a complete hostile intent story shows: (1) Higher-level goals: How the country owning the platform is motivated to attack a U.S ship. (2) Opportunity: How own ship is a logical target for attack given the country's high-level goals and any other potential targets that it could have chosen. (3) Capability: How the track is a logical choice as an attack platform given the available capabilities of the attacking country and its goals. (4) Localization: How the contact would have been able to detect own ship's location. And (5) reaching position: How the contact's actions make sense as ways of getting to an attack position quickly and safely. The story also addresses what the contact is expected to do next.

Insert Figure 3 about here

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. The causal factors at the left in Figure 1 (goals, opportunity, and capability) make up what might be called "the big picture," and they often shade the way kinematic cues are interpreted. The strongest evidence that a country's goal is to attack U.S. ships is a prior incident of doing so, and this powerfully influences the reading of subsequent events. 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, and 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. The next strongest evidence for motivation to attack is 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 can be an even subtler 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. This observation does not disprove hostile intent, but argues against it. The conflict may have a good explanation — for example, the contact did not detect or have prior intelligence regarding the command and control ship, or own ship is a more desirable target because it is an AEGIS cruiser. 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 the assessment of hostile intent (but does not prove that it exists). Capability is another useful cue in the hostile intent template. 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.

If the approaching platform is hostile, there should also be a plausible story about how it has localized own ship or is attempting to do so. This is a surprisingly frequent concern among experienced officers, and heavily influences the more standard cues provided by track kinematics. 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 the contact was too far away to have detected own ship, these cues must mean something else. Here, too, the hostile intent story can be patched up, perhaps by explaining the conflict in terms of third-party targeting support or improved equipment or training. These explanations can then be tested or evaluated. Conversely, in other incidents, a track that is too slow or too high may appear not to fit a hostile profile. However, its behavior can sometimes be explained within a hostile intent story by assuming a need

to localize the target.

The idea that stories contain characteristic events associated with intent can be generalized to other assessments besides 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. 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 engages 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. 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.

According to the R / M model, upon which this training is based, stories are needed when no pattern fits all the observations. Because they are constructed and revised through critical thinking (the STEP process), stories are not just fixed patterns or checklists of cues associated with particular intents. Each time a given type of story is used, its components are filled in differently as the decision maker searches for the most plausible explanations of conflict. It is the uniqueness of stories that makes the evaluation step so important. 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. No single element of the template is conclusive by itself. The officer must look at the "whole story" — examine all the assumptions required to make an assessment fit, and decide which story is more plausible.

Critiquing Stories

Once a complete and consistent story has been constructed, the decision maker evaluates its plausibility. One way to gauge confidence in an assessment is direct: Just ask, for example, "How confident am I that this platform intends to attack my ship (or does not intend to attack my ship)?" This approach can be seriously misleading. There is evidence that people tend to be overconfident when they provide direct estimates of this kind, even those who are (appropriately) regarded as experts in their field (Lichtenstein, Fischhoff, & Phillips, 1982). Overconfidence can cut thinking short before key issues have been explored, and it maybe one reason for unfortunate surprises or overhasty decisions. A quite different approach is to take just a few moments and *assume* that your assessment is wrong no matter how confident you are that it is true. The purpose of this exercise is not to undermine confidence. At its conclusion, you may believe your original conclusion even more strongly than before — or you may change your mind. But the best way to *earn* your confidence is to take seriously the possibility that you are wrong.

This segment of the training introduces a devil's advocate method that consists of four steps:

1. Select an important assessment, such as the intent of a contact, or an important assumption, such as that a track cannot localize own ship at its present distance.
2. Imagine that an infallible crystal ball (or some perfect intelligence source) tells you that this assessment 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 assessment 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 your 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 assessment could be wrong.)

This method is illustrated by means of several scenarios, based on interviews with officers, in which apparently compelling arguments must be questioned in order to make sense of the situation. In one such incident, an AEGIS cruiser was escorting a flagship through the Straits of Hormuz, off the Iranian coast. The cruiser detected two Iranian fighters taking off from an Iranian air base and identified them as F-4s from a brief radar transmission. Instead of heading north or south along the Iranian coast for a routine patrol, 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 turned on their search radar and kept it on continuously. The was unusual, the captain noted, because, "The Iranians did not have the maintenance capability to fly their electronics and burn them steadily." As the circles widened, the aircraft switched their radar to fire control mode, 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. This concerned the captain because, "He has to physically do that; radar didn't automatically do that."

This behavior did not fit the pattern of a routine patrol. Moreover, a disturbing number of elements for a plausible hostile intent 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 as well as the cruiser "obviously heightened our interest," according to the captain. Localization could be explained by the deliberate use of search radar despite severe maintenance problems, and locking on, while gradually widening circles brought the aircraft into position to engage. In fact, the captain noted that by illuminating own ship with fire control radar, the aircraft had already met the criterion for engagement according to the rules of engagement (ROE). Yet the kinematics of the tracks contradicted expectations. 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."

Under normal circumstances, the observation of circling aircraft in broad daylight leads to the recognitional conclusion that their intent is not hostile. In this situation, however, that recognition-based argument conflicts with other recognition-based arguments that do suggest hostile intent (e.g., locking on with fire control radar). To resolve the conflict, officers must search beyond the normal, recognitional meanings of the cues. 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 they must 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. The trainees must provide another interpretation. 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.

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). If the captain believes these aircraft to be hostile, he must be prepared to assume that at least one of these explanations, or some similar one, is the case. (By the same token, if he accepts the recognitional meaning of circling, that the aircraft are not hostile, he must assume that none of these explanations is the case.) This exercise, which often takes only a minute, can have very practical consequences. Many of the explanations can be tested; for example, those 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 may be confirmed or disconfirmed by intelligence, e.g., those regarding the aircrafts' weapons or radar characteristics. Other

explanations may be dismissed as implausible. Some can provide the basis for contingency planning in case they turn out to be true. Perhaps more importantly, knowledge of the possibilities provides the captain a real basis for evaluating the plausibility of the hostile intent assessment. In this case, the captain concluded that the intent of the aircraft was to harass rather than to attack his cruiser. Nevertheless, he developed a contingency plan for the possibility of attack, by ordering internal ship defensive systems to a high state of readiness.

In some cases, conditions under which an argument fails may be so familiar that they are recognized virtually at the same time as the argument itself. For example, in another incident 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. Flying toward U.S. ships from Iran suggested an attack; moreover, the track's altitude was too low in relation to its speed for a commercial airliner. Failure to respond to an IFF challenge further supported hostile intent. Yet the Anti-Air Warfare Officer (AAWC) had a ready 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 before flying near hostile territory and then forgot to turn them back on coming out. This non-hostile story was familiar and reasonably plausible. Moreover, the slow speed of the aircraft argued against hostile intent. Despite this tentative non-hostile assessment, as we shall see, the AAWC continued to consider the possibility that the aircraft was hostile.

In other cases, an argument may appear plausible at first blush, but has weaknesses that are not immediately recognized; they are only revealed by more deliberate critiquing. It is as if decision makers adopt a devil's advocate strategy, like the crystal ball, to elicit alternative interpretations. They assume that the argument is false — the observations are true but the conclusion is wrong — and search for an explanation. This is precisely what the AAWC did to take seriously the possibility that the low-flying non-squawker was hostile. Slow speed conflicted with expectations regarding an attacking aircraft. But 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"). 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 disconfirmed, since the aircraft was not flying directly toward own ship. The AAWC then continued to assume hostile intent and forced himself to find still another explanation for slow speed: 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. 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 is a last-ditch effort to save the assumption that the aircraft is hostile. The explanation 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.

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 carried out 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.

After about 30 minutes of practice, trainees start considering exceptions to conclusions spontaneously, without explicitly invoking the device of a crystal ball. Assessments or plans that survive this questioning stand a better chance of being true and successful.

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. 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. For these officers time for thinking or observing has run out as soon as own ship is within range of the contact's weapons. At this point, these officers turn off their brains and are ready to shoot. 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.

Stakes of an error. 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.

Situation typicality. Less experienced officers learn familiar, highly repetitive patterns such as

commercial air schedules, corridors, and speeds, or routine patrol routes, speeds, and altitudes. They are very good at detecting non-hostile behavior that fits such patterns, and as good or better than experienced officers in detecting deviations from them, especially if the contact is heading towards own ship. Unfortunately, this is where many inexperienced officers stop. More experienced officers are more sensitive to the fact 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, even if it also matches a few aspects of an attack pattern (e.g., heading and speed). They are more likely 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). Thus, while both sets of officers notice departures from stereotypical friendly or neutral patterns, the more experienced officers also recognize departures from hostile patterns. In sum, more experienced decision makers realized that a situation was ambiguous in cases where less experienced decision makers did not. As a result, they were more likely to see the need for stories and the iterative processes by which they are constructed and evaluated.

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 of participants available for testing and the time and quality of training. On the one hand, study 1 had more participants; as a result, we utilized a control group as well as a pretest-posttest comparison. Study 2 had fewer participants, and only a pretest-posttest comparison was utilized. On the other hand, training conditions in Study 2 were far superior to those in Study 1. In Study 1, only 90 minutes were available for training, and the entire test procedure (familiarization, pretest, treatment, and posttest) was compressed into a single long day (9 hours). Moreover, practice in study 1 utilized paper-and-pencil examples (although testing did involve an automated simulation). In study 2, four hours were

available for training, the procedure was spread across 5 two-hour sessions, and an interactive simulation was used for more realistic practice during training.

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 many cases, these modifications replicated the situations that had been described by officers in interviews. Participants performed in these scenarios individually, acting as TAOs, followed by feedback from the instructor and 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 of 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 assessments of the intent of the designated track and estimate confidence in each of those assessments. (3) Select an assessment of the designated track that the participant did not agree with and defend it. (4) Identify evidence that conflicted with an (experimenter-designated) intent assessment and then defend the designated assessment. (5) Describe actions the participant would take at this time in the scenario. Measures based on these questions covered every phase of the STEP process: Stories — the number and variety of issues considered when evaluating an assessment; Test — the amount of conflicting evidence that a participant identified, and the number of explanations generated to patch up stories; Evaluate — the number of alternative assessments generated and the accuracy of the assessment favored by the participant; and Plan — the frequency with which contingency plans were created in case an assessment was wrong. 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.

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.

In 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 ($\chi_1^2 = 10.816$, $p = .001$). 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. Apparently 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. The ability to generate alternative assessments of a track helps officers evaluate their favored assessment. By suggesting alternative interpretations of observations that seem to support the favored assessment, it exposes hidden assumptions in the current story. Such assumptions 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; once begun, they can stop at any time, if circumstances have evolved, and act on their best current assessment. 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, presumably, 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 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. We would expect consensus to grow if training helped officers to converge on a correct assessment. 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 scenario of Study 1 that elicited training effects on assessment quality, the intent of the approaching track could have been participation in an on-going search and rescue (SAR). However, it could also have been to use the SAR operation as 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 of training by officers in study 2 was also 3.7. Officers with some tactical experience in the Navy or other military services tended to rate the training positively than were 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. 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

The research described here has moved from the empirical analysis of decision-making incidents to the development of a naturalistic model of uncertainty handling. From there, it has moved to the development of a training strategy based on the model, and to successful initial testing of that training strategy. Along the way, a variety of lessons have been learned regarding both theory and practice, and the link between them (Salas, Cannon-Bowers, & Blickensderfer, 1997):

1. There is evidence for a set of critical thinking skills in proficient tactical decision making. 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.

2. A plausible 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. Cognitive task analysis is used to identify the content of the training in a particular domain. Simulation, feedback, and performance measures are used to support practice.

3. Critical thinking processes can be effectively taught by means of this strategy. 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%.

3. There is also evidence that the critical thinking training strategy can improve outcomes, i.e., the accuracy of assessments. Agreement 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.

4. Critical thinking training for teams will not simply replicate the training strategy we have developed for individuals, but will introduce new processes and skills. For example, in a team training strategy that we are now testing at SWOS, officers are taught to verbalize their assessments in periodic situation updates (Entin, Serfaty, & Deckert, 1994). When the stakes are high and time is available, these updates also mention problems with the assessments, such as missing, unreliable, or conflicting evidence. Such communications may increase the ability of other officers to provide needed information or insights.

Officers are also being taught to work together as devil's advocates to generate new interpretations of evidence and alternative assessments.

Critical thinking skills play an important role whenever decisions must be made under uncertainty and time pressure. Training in such skills may be a valuable adjunct to training in a variety of military and non-military contexts.

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. (1997). 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., & 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.

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. 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.

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

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.

Nelson, T. O., & Narens, L., (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), Metacognition (pp. 1-25). Cambridge, MA: The MIT 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.

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	60	35
Design	Trained (40) versus Control (20) x pretest-posttest	Pretest-posttest
Duration of training	90 minutes	4 hours
Practice tools	Paper and pencil	Simulation (DEFTT)
Scheduling	1 8-hour session	5 2-hour sessions

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 in regard to assessment	7% improvement	30% improvement
Number of conflicting pieces of evidence identified	52% improvement	58% improvement
Number of explanations of conflict generated	26% improvement	27% improvement
Number of alternative assessments generated	10% improvement	41% improvement
Accuracy of assessment	42% improvement	18% improvement
Agreement on assessment	14% improvement	41% improvement
Confidence in assessment	13% increase	20% increase
Frequency of contingency planning	217% improvement	[not analyzed]

Subjective evaluations of training	73% positive	71% positive

Figure Captions

Figure 1. Components of the Recognition / Metacognition model.

Figure 2. The STEP process for critical thinking.

Figure 3. Hostile intent story template.

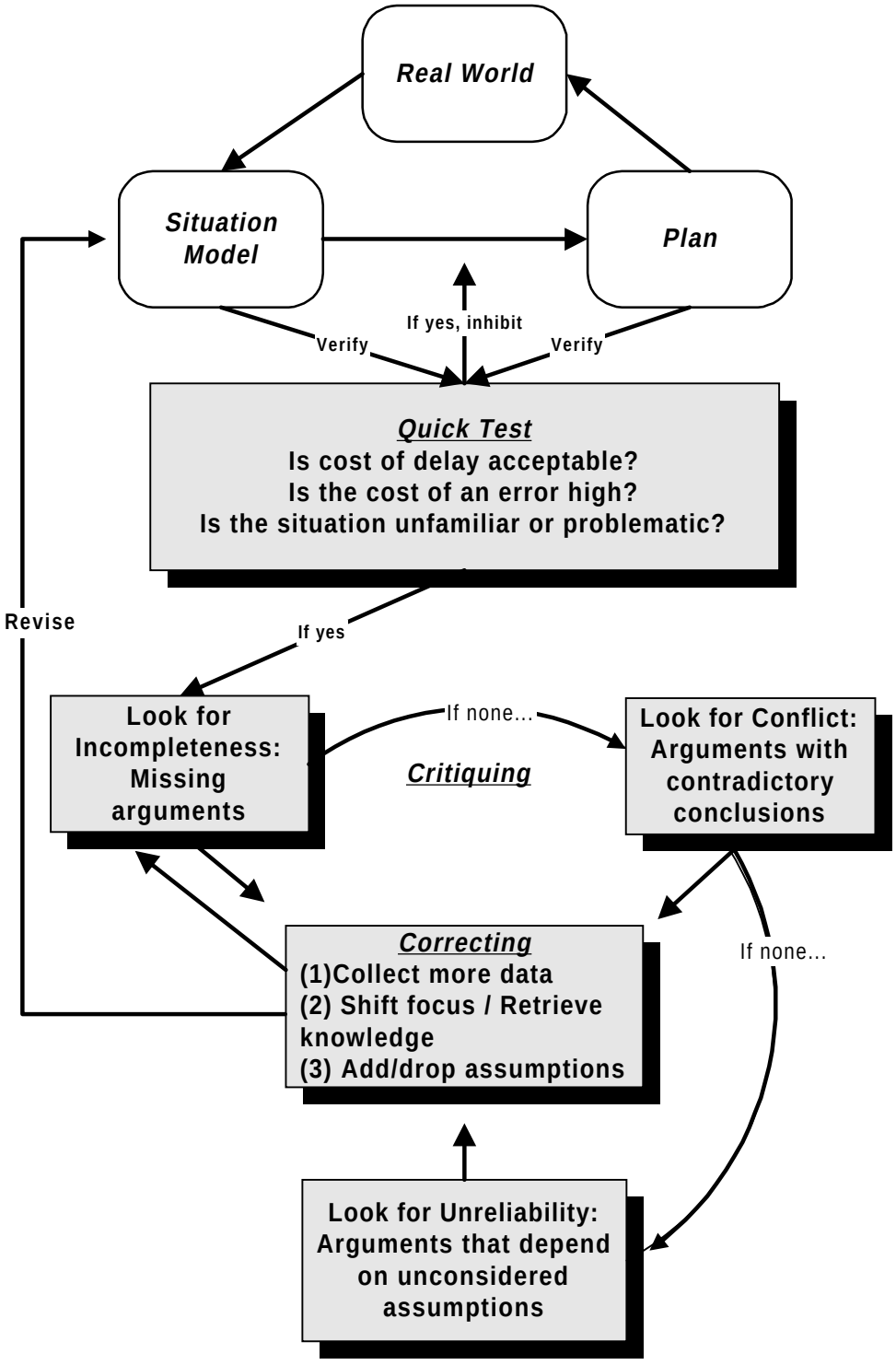
Real World
Situation Model
Plan
Verify
If yes, inhibit
Verify
Quick Test
Is cost of delay acceptable?
Is the cost of an error high?
Is the situation unfamiliar or problematic?
Revise
If yes
Look for Incompleteness: Missing arguments
If none...
Critiquing
Look for Conflict: Arguments with contradictory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

Figure 1

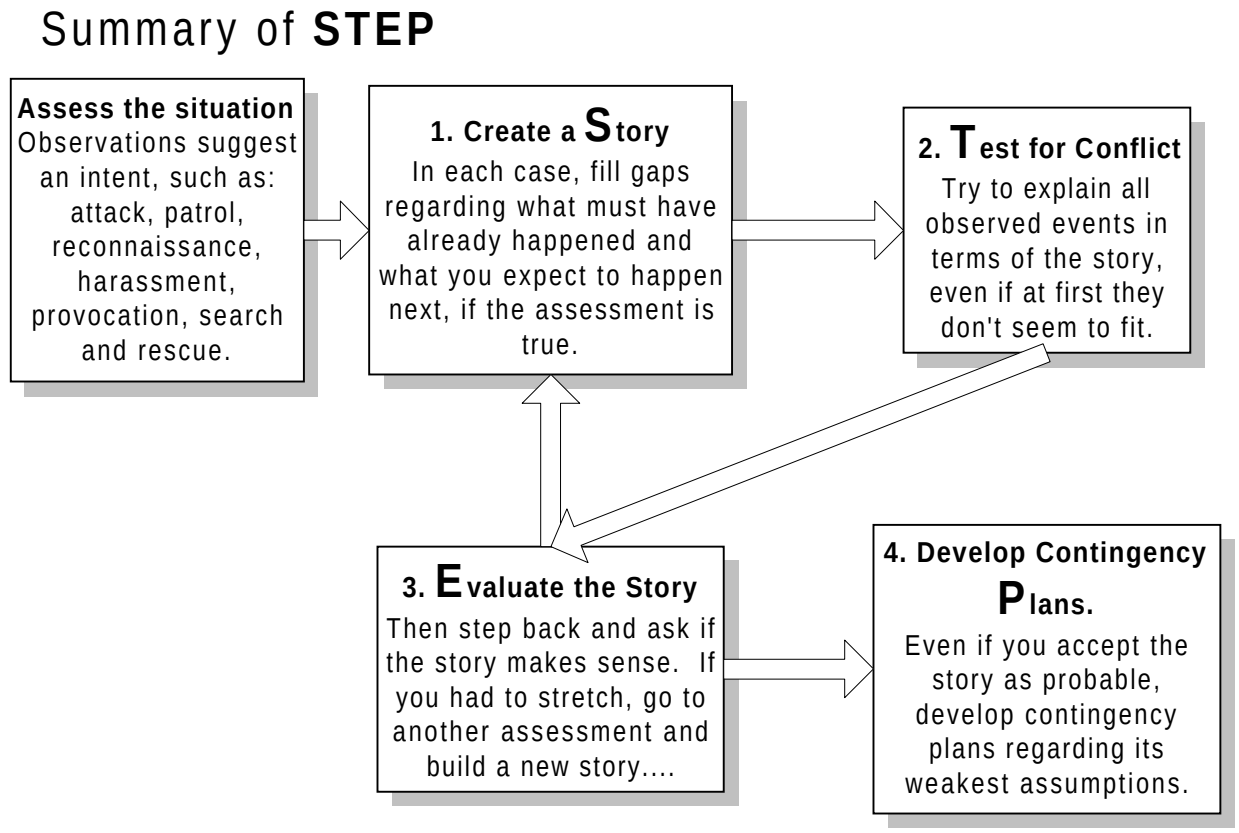
Summary of STEP
Assess the situation
Observations suggest an intent, such as: attack, patrol, reconnaissance, harassment, provocation, search and rescue.
1. Create a Story
In each case, fill gaps regarding what must have already happened and what you expect to happen next, if the assessment is true.
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

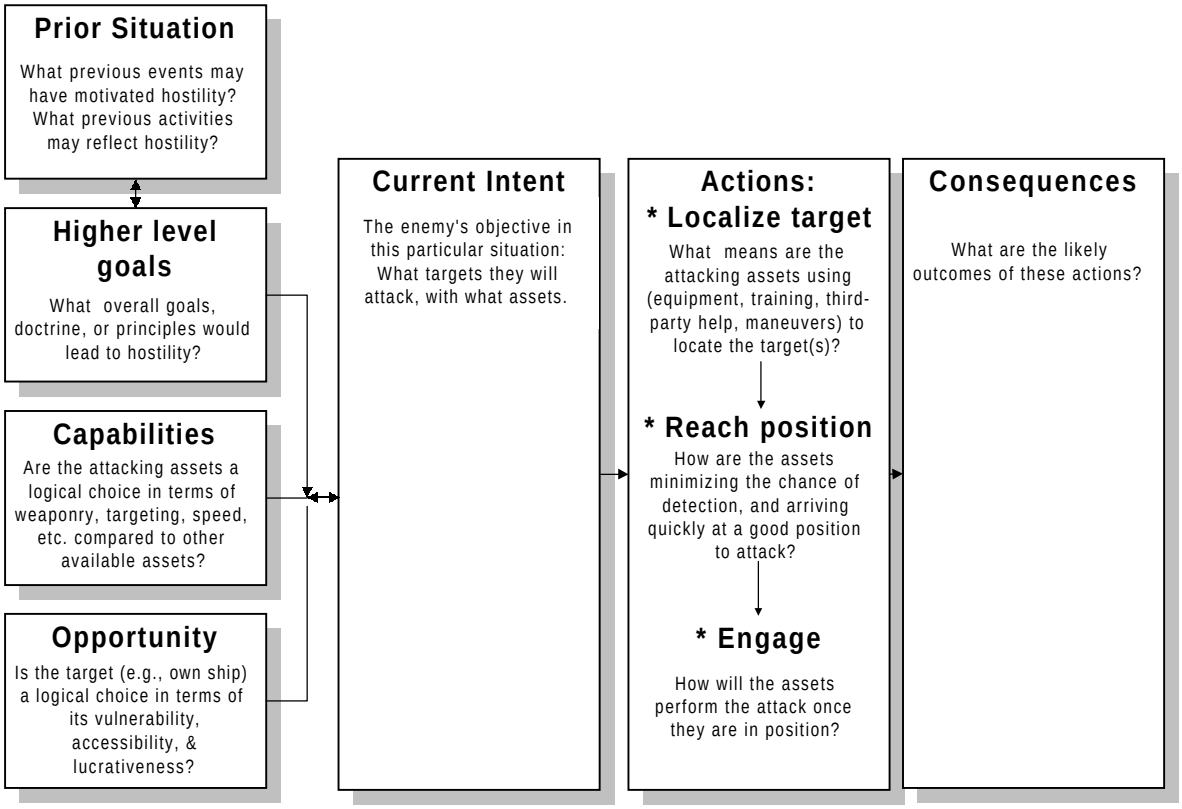
Prior Situation
What previous events may have motivated hostility? What previous activities may reflect hostility?
Higher level goals
What overall goals, doctrine, or principles would lead to hostility?
Capabilities
Are the attacking assets a logical choice in terms of weaponry, targeting, speed, etc. 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