A COGNITIVE BASIS FOR AUTOMATED TARGET RECOGNITION INTERFACE DESIGN

Prepared by:

Marvin S. Cohen and Martin A. Tolcott

Cognitive Technologies, Inc.
4200 Lorcom Lane
Arlington, VA 22207
(703) 524-4331

Prepared for:

Director, U.S. Army Laboratory Command
Human Engineering Laboratory
ATTN: SLCHE-AD/Mr. Harper
Aberdeen Proving Ground, MD 21005-5001

FINAL TECHNICAL REPORT

15 December 1992

ABSTRACT

The successful introduction of automated target recognition systems into combat environments will depend on how well they interface with the knowledge and processing strategies utilized by human operators. The goal of the present research is to investigate cognitive issues in human recognition performance, and to explore their implications for human interaction with automated recognition devices. Theoretical background, specific hypotheses, and a plan of experimental research are laid out in four areas:

1. Stages of visual processing and implications for the display of visual and non-visual data
2. Favored levels of generality/specificity in verbal categorization and implications for the display of ATR classification conclusions
3. Reasoning with mental models and implications for the display of uncertain ATR conclusions
4. Strategies for verifying recognitional conclusions and implications for the display of prompts and data under uncertainty

The products of the planned research will include empirically supported general design principles for ATR-human interfaces, and a set of specific ATR interface designs.

TABLE OF CONTENTS

1.0 INTRODUCTION

1.1 The Problem

The pilot of an attack helicopter may emerge above the trees or hills for only a minute or two to assess the battlefield situation before remasking. During that brief period, he must collect accurate and relevant information about potential targets while minimizing his own exposure to attack. The pilot must be able to detect and discriminate friends and foes in a highly target-dense, rapidly changing, and visually and electronically noisy environment; he must classify targets that are relevant to his mission (e.g., tanks versus armored personnel carriers versus anti-air artillery), and prioritize them for attack. After remasking behind trees or terrain, he must decide whether to unmask again in a different location to collect more information, or to pop up to engage his target.

Target recognition has become a crucible of success on the battlefield not only for helicopters, but for virtually every weapons platform. The reason for its importance lies both in the recent evolution of U.S. war-fighting doctrine and in the development of new sensor and weapon technologies. Exploitation of U.S. night-fighting capabilities, for example, degrades the quality of optical information available for recognition decisions by both helicopters and tanks. Utilization of stealth technology and rapid maneuver tactics constrains the information obtainable from communications and from active sensors like radar (which would alert the enemy to one's own presence), while compressing the time in which recognition decisions must be made. At the same time, increased enemy mobility and speed and improved enemy sensor and weapon ranges reduce the time available for recognition; and information denial techniques (such as camouflage, stealth, electronic countermeasures, and tactical deception) increase the uncertainty that such decisions must resolve. Nine U.S. soldiers and nine British soldiers were mistakenly killed by U.S. aircraft during Operation Desert Storm. But accurate target identification had become a major U.S. military concern well before the Persian Gulf War (see, for example, *Defense News*, March 25, 1991), and it is easy to imagine scenarios in which it would have played a much more crucial role in determining the success of battle.

1.2 Current Approaches

One approach to target recognition is the development of cooperative identification systems, e.g., distinctive visual markers for friendly vehicles, or systems for the electronic exchange of codes among friendly aircraft and vehicles. The limitations of these devices are obvious: In a sophisticated battlefield, they both expose friendly units to detection and identification by the enemy, and may be exploited by enemy units who mimic the friendly codes. Current research and development interest centers primarily on so-called non-cooperative target recognition (NCTR) systems.

It is natural to think of non-cooperative target recognition as primarily a sensory or perceptual problem. From this point of view, improvement in recognition accuracy and speed will come by providing more and better target information to the human operator: i.e., (a) improved sensors and analyzers (e.g., infrared, electro-optical image enhancement, synthetic aperture radar, laser radar, and others), and (b) improved cockpit display technologies, e.g., higher resolution, color, digital interactive displays with symbolic overlays, and so on. Artificial intelligence work on the user-computer interface promises even more dramatic input/output technology for the future: e.g., spatial data management, natural language understanding, voice I/O, object-oriented direct-manipulation interfaces, and multi-media three-dimensional virtual environments. The downside of this approach (taken by itself) is that it may leave the operator overloaded in environments where the sheer number of targets overwhelms his ability to detect and recognize them.

A third approach to target recognition is to automate the human's role in interpreting the sensor outputs. Automated Target Recognition (ATR) devices are under development which aim to automatically detect, track, classify, and prioritize objects of interest in a sensor image. In many cases the goal of such systems is to reduce the role of the human to that of a passive observer (Toms and Kuperman, 1991). Despite considerable progress, the performance of ATR systems has not yet been sufficient to achieve this goal. Current systems often fail under novel conditions (e.g., of weather, time of day, target aspect, or target configuration), under degraded observation conditions (e.g., low contrast or high clutter), or in the face of enemy countermeasures. In order to reduce the number of missed targets to an acceptable level, false alarm rates must often be set intolerably high, once again overloading the human operator.

A final approach regards ATR systems and human operators as partners. At least in the near future, humans will fill the gaps in ATR performance and ATR systems will relieve human workload. From this point of view, target recognition is only one phase or component of the human's overall task: For example, he must also decide how long to remain exposed, where to look and with what sensors, whether to look again, when to fire, where to fire, and with what weapons. Designs for human-ATR interaction must take this entire complex of activities into account. It is to this approach that we now turn.

1. 3 A Cognitive Approach

When ATR systems are initially fielded, they will almost certainly require human partners. Even in the longer term, human-ATR collaboration may be beneficial, providing flexibility and robustness not available in a fully automated system. Different issues may come to the fore as ATR systems are more thoroughly automated over time. In the short term, the emphasis will be on promoting human intervention in the recognition of selected individual targets: e.g.,

- methods for directing user attention to targets where ATR conclusions are uncertain,
- natural and effective displays of sensor data as well as ATR conclusions for those targets,
- simple methods for providing human inputs.

In the longer term, the human role may shift from individual targets to a more executive-level function of monitoring and fine-tuning the operation of the ATR system as a whole. At this stage, the most pressing human-computer interaction issues may be:

- alerting users when ATR data, assumptions, or methods may be faulty,
- providing effective explanations of ATR reasoning and results, and
- permitting human operators to override or adjust internal ATR parameters or rules.

Despite the importance of user-ATR collaboration, there has been virtually no research on the *cognitive* impact of ATR devices. Very little research has addressed the decision making skills of the user: How can computer-implemented analysis and transformation of sensor data be optimally interfaced with human knowledge structures and recognition strategies, in the context of the pilot's battlefield missions?

Automation without concern for human cognitive constraints may disrupt rather than support the human's role in the recognition process. For example, there is often a mismatch between the user's way of thinking and that imposed on him by the computer in uncertainty handling. There is ample evidence that people do not engage in the exhaustive generation of hypotheses, or the precise quantification of the strength of evidence, that are required by conventional models. Displays of multiple alternative possible classifications along with numerical uncertainty by automated devices may therefore be both meaningless and distracting to human operators. On the other hand, without some indication of uncertainty, automated displays may encourage overconfidence in users and fail to direct their attention to problematic results. In other cases, operators may be underconfident, exposing themselves to unnecessary risk while collecting superfluous information.

Another problem is the transition from attention to individual targets to a more supervisory role by human operators. Such a transition may never be complete. Yet displays that optimize overall system monitoring may degrade human recognition performance with individual targets; the need to fine-tune system parameters may clash with the time constraints of recognizing and engaging individual targets; prioritization of targets for engagement may require transparent displays that enable users to concretely visualize an evolving tactical battlefield undistracted by *either* the fine details of specific targets *or* information about how the system processes information. Unless the mix of operator roles in future systems is carefully defined and supported, there is danger that human operators may sit passively in front of automated consoles that operate too quickly and too incomprehensibly for humans to make any effective contribution at all.

The present research focuses on the development of cognitive collaboration between ATR systems and human operators. Target recognition interweaves perceptual and cognitive components, and there is often no hard and fast separation between the two. But there are at least four areas in which cognitive questions seem crucial:

1. *Visual processing:* How do operators represent and process visual data? When and how do they decompose images into parts? How do they use spatial and temporal context (such as convoys or typical SAM deployment patterns), or self-generated templates to classify targets? How do they combine image data with non-visual data? What are the implications for how ATR data should be displayed?
2. *Verbal categorization:* What determines the label that operators apply to a target, i.e., the level of generality or specificity at which they choose to describe it? How do operators represent the patterns of similarity and dissimilarity among different classes of targets and non-targets? How do different levels of categorization reflect this structure? What are the effects of typicality and goals? What are the implications for the way ATR conclusions should be reported?
3. *Uncertainty handling:* How do operators represent uncertain conclusions? When do they attempt to resolve the uncertainty by collecting more data? When do they simply accept it? When do they resolve it by making assumptions? What determines the assumptions they choose (e.g., worst-case, best-case)? What are the implications for the way ATR's should report uncertain results?
4. *Strategic control of processing and metacognition:* How do operators allocate their attention across a display? How do they evaluate their confidence in automated results? How do they evaluate confidence in their own conclusions? How are decisions about collecting more information influenced by mission constraints? What are the implications for the allocation of effort between ATR devices and human users?

Current approaches to decision aiding have typically offered either too much help or too little; they have emphasized either adaptation to the problem (i.e., compatibility with "correct" or "optimal" decision models or expert rules) or adaptation to the user (i.e., providing tools and displays at the user's request with little guidance as to how to use them). Less attention has been paid to the challenge of simultaneously providing flexibility and guidance, i.e., adapting to the user *and* to the problem. As a result, aids that truly integrate human and computer decision processes have evolved (if at all) by trial and error. It should not be surprising, under these circumstances, if users become both less confident in the system's output and less able to utilize their own knowledge and abilities. A result may be the drift toward a level of complete automation in which the human has virtually no role at all.

1.4 Phase I Objectives

This report summarizes the results of Phase I of a Small Business Innovative Research contract on Human Performance Issues in Automatic Target Recognition. Our focus is on the cognitive aspects of human target recognition performance, and enhancement of the human contribution to so-called automatic target recognition (ATR) systems.

This research reflects two basic principles: (i) that human participation in automatic target recognition processes can enhance overall system performance, and (ii) that the design of effective *interactive* recognition systems must be based on appropriate research from cognitive science: in knowledge representation and categorization, the executive control of cognition, and human decision errors. The overall objective of Phase I is to investigate the feasibility of experimental research that tests these principles.

Our more specific Phase I goals are: (1) to develop hypotheses about recognition performance in an ATR context based on the cognitive science literature, (2) to develop interactive display concepts based on those hypotheses, and (3) to suggest an experimental plan that tests both the hypotheses and the display concepts. The experiments themselves will be carried out in Phase II of the project.

To become familiar with the ATR environment, we have met with researchers, system designers, and with potential ATR users:

- Researchers (and the project's sponsors) at the U.S. Army Human Engineering Laboratory, Aberdeen Proving Ground, Maryland;
- Researchers working on human factors, ATR test and evaluation, and multi-sensor fusion at the U.S. Army Night Vision and Electro-Optics Directorate, Fort Belvoir, Virginia;
- Engineers working on multi-sensor fusion systems at the Westinghouse Electric Corporation, Electronic Systems Group, Baltimore, Maryland; and
- Active-duty helicopter pilots (Apache and Delta) in the U.S. Army 18th Airborne Corps, Fort Bragg, North Carolina.

1.5 Overview

In Section 2.0 we describe our basic approach to decision aiding and interface design. This approach is both *Personalized and Prescriptive*: We emphasize adaptation to an individual user's cognitive capabilities and knowledge, while at the same time guarding against errors or pitfalls to which his preferred approach may lead.

In Section 3.0 we apply the Personalized and Prescriptive Aiding framework to ATR interface design. We (a) provide a psychological framework for modeling human target recognition performance, (b) derive specific hypotheses about human target recognition performance based on that framework, and (c) describe ATR interface concepts based on those hypotheses. The design concepts cover four areas:

- the display of image data
- the display of ATR conclusions,
- the display of uncertain conclusions, and
- strategies for operator-ATR task-sharing, especially user verification of ATR results

Section 4.0 outlines an experimental plan for testing both the performance hypotheses and the display concepts described in Section 3.0.

2.0 PERSONALIZED AND PRESCRIPTIVE AIDING

Design of displays to support decision making has often veered between two extremes: *technology-driven* and *status-quo-driven*. In the first case, the problem-solving strategy is determined by a technology, such as mathematical optimization, decision analysis, or a favored artificial intelligence reasoning method, without regard for the user. In the second case, the users' current methods are simply automated without regard to whether they are the best way to solve the problem.

An alternative approach to display design, called Personalized and Prescriptive Aiding, is jointly *user-driven* and *problem-driven*. The designer's goal is (1) to adapt the system to the cognitive capabilities and preferences of the user, without necessarily duplicating the status quo, and (2) to solve the problem effectively, but without excluding the user's potential contribution (Cohen, Leddo, and Tolcott, 1989; Cohen, in press).

Personalized and Prescriptive Aiding involves the following components:

- Understanding the decision making processes and strategies of potential users, including individual differences. This may involve interviews, observation of performance, simulator tests, or formal experiments.
- Understanding the problem and effective methods for solving it. This may involve mathematical modeling, elicitation of expert knowledge, or a combination of expert judgment and analysis, depending on the problem.
- Comparison of user decision making strategies with effective methods, to identify the strengths and weaknesses of different user strategies.
- Design of displays and user-system dialogues that facilitate the strengths and guard against the weaknesses of a user's approach. Such systems support user-preferred problem-solving strategies, while providing prompts and other safeguards against potential pitfalls associated with those strategies. The result should be a system which performs better than either the user or a completely automated system.

In Section 3.0 we will explore the application of these methods to the design of interactive ATR displays. In the remainder of this section, we discuss some of issues pertaining to personalization and to prescriptive support, respectively.

2.1 Personalization and Adaptive Flexibility.

There is evidence in a variety of contexts that people use different decision-making strategies in different tasks and at different stages in the same task (e.g., Payne, 1976; Wright and Barbour, 1977; Svenson, 1979; Russo and Dosher, 1981). A strategy is any consistent pattern of actions in response to perceived conditions, where the actions may include information gathering and information evaluation, and the conditions may include the results of previous actions. For example, in the face of uncertainty, some decision makers prefer to evaluate actions with respect to the worst-case possibility, while others prefer to evaluate options against average or expected outcomes (Cohen, 1987).

There is little evidence that the same individual is consistent across tasks in the decision making strategies that he employs (e.g., Libby and Lewis, 1977; Sage, 1981; Hammond et al., 1984). Decision makers may well differ in such cognitive styles as "intuitive" and "analytic," but there has not yet been a reliable mapping of such coarse traits onto consistent patterns of information input and output (Huber, 1982). Many variations in strategy are determined by features of the *task,* rather than the *individual*. For example, decision makers are more likely to screen options with respect to cutoffs or goals when there are a large number of options to consider, but are more likely to look at tradeoffs among goals when there are only a few options (Payne, 1976). Hammond and his group have shown how differences in the way information is presented can lead (in a very reasonable way) to processing strategies that appear intuitive or analytic.

Huber (1982) argued that a system which attempts to infer and model enduring cognitive characteristics of a user and impose an appropriate decision making strategy on him is both impracticable and undesirable. It is not likely to be very successful in predicting a consistent strategy across widely varying tasks and situations. And when it is wrong, user frustration is likely to be high, even if the user can override the system's choice. Lehner, Mullin, and Cohen (1988) point out the difficulty users face when the system unpredictably (and unjustifiably) shifts display or processing modes.

A more promising avenue to explore is a system which can flexibly and quickly accommodate user strategies *for adapting to different kinds of tasks*.

Flexibility to accommodate individual differences is not sufficient in and of itself. A vast number of logically possible strategies could be made equally easy. But such a system is more likely to overwhelm users with too many choices than it is to facilitate performance. A more reasonable objective is to identify a relatively small subset of strategies (like worst case versus expected case evaluation, or evaluation with cutoffs versus tradeoffs) that are (a) most often utilized by decision makers in a particular domain and (b) which appear to be adaptive. Decision aids may then be *tuned* to facilitate selection by the user from this smaller group of strategies. Such aids are *adaptively flexible* in their responsiveness to likely user needs and task demands. They adapt to the problem at the same time as they adapt to the user.

The "user model" (Lehner, et al., 1987) that is incorporated within such a system at any given time consists of (a) its knowledge of the decision strategy currently being employed by the user, (b) knowledge of how the user has executed the strategy and the results of applying it so far, and (c) a set of adaptive interface functions which channel and/or prompt the user to undertake appropriate new actions within the strategy. Knowledge of the currently employed strategy is obtained either explicitly by direct user choice or implicitly by system observation of the user's pattern of interaction with the system. Neither (a), (b), nor (c) is very expensive computationally, and none of them needs to involve "artificial intelligence" in any very sophisticated sense.

2.2 Cognitive Interface Dimensions

A small set of cognitive interface dimensions can be used to define (at a very general level) the strategies that decision makers can adopt while interacting with a computerized aid. Each of these interface dimensions can be personalized to accommodate user strategies. Generic dimensions of personalization that pertain to object recognition, together with examples from the ATR context, are the following:

- *Level of aggregation of displayed information:* Should relatively raw data be displayed (e.g., video inserts, or chips, of target of interest), intermediate conclusions (e.g., labeled feature templates showing the ATR's perceptual analysis of the object), and/or final conclusions (e.g., a classification of the object as a tank)?
- *Generality versus specificity of conclusions:* Should a target be labeled by the ATR as a T672, a tank, a tracked vehicle, or merely as an object of interest?
- *Amount of supporting data:* What is the spatial and temporal extent, size and resolution, of the displayed data? Should all evidence for and against the conclusion be displayed, only the most crucial evidence, only evidence that confirms (or disconfirms) the recommended conclusion, or only the crucial evidence within certain spatial, temporal, or resolution-based constraints?
- *Alternative possible conclusions:* Should all possible conclusions be displayed (e.g., the target may be a truck or a tank), conclusions above a certain probability, only the single most likely conclusion, or only the worst- (or best-) case conclusion?
- *Selection of objects:* How should the user's and the system's attention be directed? Should objects be selected based on proximity to own platform, threat to own platform, mission tasks, inconclusiveness of ATR analysis, inconclusiveness of user analysis, or some combination of the above?

These dimensions pertain not only to system *displays* but also to user *inputs*. Individuals differ in the information or decisions that they prefer to provide, have time to provide, or feel confident providing, e.g., in level of generality/specificity, level of aggregation, amount and type of information, and precision of judgment.

We turn now to ways in which decision aids may complement and improve user strategies, rather than simply adapting to them.

2.3 Prescriptive Aiding: Channeling and Advisory Prompting.

Decision aids that are tailored exclusively to a particular decision maker's preferred strategy may miss opportunities to improve decisions. Traditional decision aids, on the other hand, often require users to adopt radically different techniques of problem solving (e.g., based on normative models or strategies preferred by designers or other users). In doing so, they may throw out the baby - user knowledge and experience - with the bath water - user errors or inefficiency. Personalized and prescriptive decision aiding attempts a less radical, more "surgical", correction of user errors and inefficiencies. It facilitates the user's *overall* approach to the problem, when it appears basically reasonable, but guards against *specifically identified* shortcomings. It promotes relatively minor amendments to preferred approaches that may produce significant increases in their efficiency and accuracy.

Two types of prescriptive methods are utilized: channeling and advisory prompting. The difference between them is largely one of tactics:

- *Channeling* is implicit and proactive. It encourages users, in advance, to adopt variants of their own preferred strategies that are less susceptible to errors, by structuring the display of information in such a way that those variants become natural and simple to execute. Channeling may be appropriate when prior research suggests that errors commonly result from certain strategies (e.g., ways of representing a problem, generating options, assessing uncertainty, evaluating outcomes, or making choices). Channeling can be built into the dialogues corresponding to those strategies, to increase the chance that the user will adopt amended versions of the strategies that avoid all or most of the errors.
- *Advisory prompting* is explicit and reactive. Advisory prompting is based on real-time monitoring of the user's actions, projection of their possible outcomes, and comparison with the system's model. A prompt occurs if the discrepancy is above a threshold of significance, i.e., if the user's strategy appears to the system to be leading to significantly suboptimal choices. The system may also monitor tasks performed by the computer and prompt where a human contribution might improve results. Thus, in advisory prompting the computer senses weaknesses in a strategy, whether its own or the user's, and offers (or requests) help.

Channeling occurs, for example, when a menu announces the availability of information about X (e.g., imagery based on other sensors) in a situation where users may tend to request only Y (e.g., imagery based on only a single sensor); or when information about X (e.g., the most likely identification) is automatically provided whenever Y (e.g., the worst-case identification) is requested.

Advisory prompting occurs, for example, when the user is alerted that information he has not requested to see (e.g., imagery from a second sensor) appears *on this occasion* to lead to significantly different conclusions from information he has requested to see (e.g., imagery from a single sensor).

Figure 1a summarizes the rationale for Personalized and Prescriptive aiding. User-preferred strategies may reflect experience and knowledge that is not captured in automated analytical models of a problem; analytical models may incorporate efficiencies and avoid errors in a way that is not possible in human approaches.

Figure 1b illustrates the complementarity between automated model and user strategy that Personalized and Prescriptive aiding tries to bring about. The goal of personalization is to effectively tap user experience and capabilities. The goal of channeling and prompting is to foster variants of the user's preferred approach that

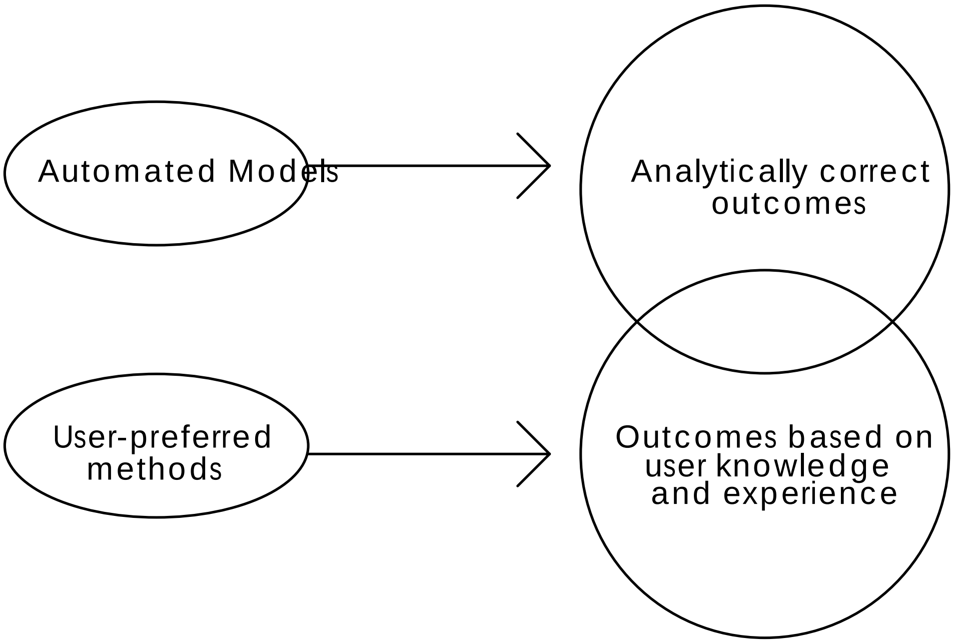
Automated Models
Analytically correct outcomes
User-preferred methods
Outcomes based on user knowledge and experience

Figure 1a. Analytical methods and user strategies may have complementary strengths and weaknesses.

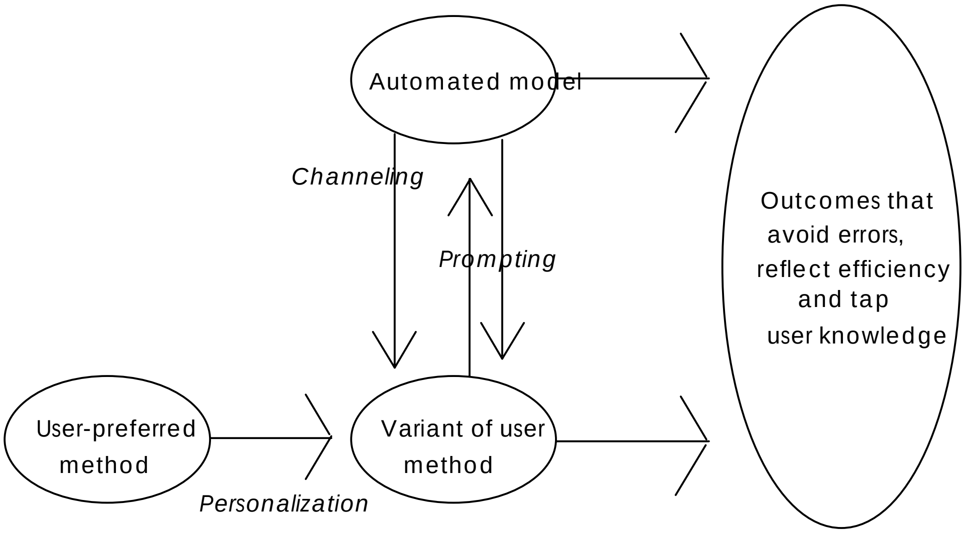

Figure 1b. Collaboration between automated model and user method, through personalization, channeling, and prompting.

correct the specific shortcomings while retaining the advantages (in terms of knowledge and experience) of the preferred approach.

The balance between personalized and prescriptive components varies with the application and, for a given application, with the context in which it is used. Essentially, prescription is more dominant in high workload situations where the computer takes the lead. Personalization is predominant in high stakes, low workload situations, where the human takes the lead. We turn now to some implications of Personalized and Prescriptive Aiding for automation and the allocation of cognitive tasks between users and computers.

2.4 Allocation of Cognitive Tasks and Subtasks

Traditionally, task allocation in human-machine systems has been according to the purported strengths of each (e.g., Fitts, 1951). Such methods have for the most part produced a fixed allocation of broadly defined activities, e.g., assigning numerical computation and long-term data storage to the computer and option generation to the human. As noted in Chinnis, Cohen, and Bresnick (1984) and Cohen, Brown, Seaver, and Ulvila (1982), this approach fails in application to computer-assisted reasoning. First, it is not fine-grained and flexible enough: rapid variations in task demands and in decision-maker expertise from task to task, or subtask to subtask, are not captured. Secondly, it is too machine-oriented: resulting task assignments may not form a meaningful or organizationally acceptable pattern for human users; users may be unprepared to take over in case of machine dysfunction. Finally, it is too conservative: novel approaches to human-computer collaboration (e.g., personalization, channeling and prompting) may be overlooked.

If static generalizations regarding human-machine superiority are inadequate, what sort of guidelines for cognitive task allocation might take their place? An important first step is to make a distinction (e.g., Sheridan, 1987) between (a) primary responsibility for the performance of a task or subtask and (b) monitoring that performance and intervening in case of trouble. For example, Wiener (1988), in an aviation context, proposes a matrix in which primary performance of a task (which he calls "controlling") may be either manual or automated, and in which the monitoring function may also be either manual or automated. If both task performance and monitoring are automated, the operator is likely to suffer boredom, complacency, and erosion of confidence. If both are manual, on the other hand, there will be fatigue, high workload, and failure to detect critical conditions. Wiener thus proposes two primary modes of task allocation: (1) operator performing and computer monitoring; and (2) computer performing and operator monitoring. This scheme, however, does not address three questions:

- When each mode is appropriate, and who should decide.
- What the computer should be monitoring when the user is in control of the task.
- What the pilot should be monitoring when the computer is in control of the task.

We will discuss the first question in Section 2.4.1, and the second and third questions in Section 2.4.2.

2.4.1 Control over task allocation

A simple hypothetical example may shed some light on control over task allocation. Suppose that a given task (e.g., target recognition) can be performed under a variety of conditions that affect the relative advantage of the operator and the automated device. Such conditions might include type of target, distance of the target, weather conditions, presence of countermeasures, and so on. In table 1, columns 1 and 2 assume that the relative expertise of the user and the aid on a particular task or subtask depends on which of six different (equiprobable) situations is the case. In some situations (A, B, F) the user is better; in others (C, D, E) the aid is better. Unless some complementarity of this sort exists, there is no point in introducing cooperative processing between user and machine at all.

Table 1. A hypothetical task allocation problem.

PERCENT CORRECT

	1	**2**	**3**	**4**	**5**	**6**
Situation	**User alone**	**Aid alone**	**Ideal allocation**	**Random allocation**	**Ideal allocation + 50% error**	**Random alloc. + 50% error**
A	90	40	90	65	45	42.5
B	80	60	80	70	40	50
C	20	90	90	55	90	50
D	60	70	70	65	70	50
E	50	80	80	65	80	52.5
F	40	30	40	35	20	25
Overall	56.7	61.7	75	59.2	57.5	45

Notes: Overall percent correct (last row) assumes situations are equally likely.
Random allocation assumes user and aid perform task 50% of the time in each situation.
"50% error" in columns 5 and 6 assumes user accuracy is 50% of accuracy in column 1.

When the user performs the task manually in all situations, the percentage of correct answers is 56.7% (column 1). When the task is always performed automatically, the percent correct is 61.7% (column 2). While the aid is, on average, better than the user, there is a significant opportunity for synergy: An ideal allocation of the task according to relative expertise in each situation yields a percent correct of 75.0% (column 3). This is considerably better than the 61.7% achievable by the aid alone.

Suppose the user must determine which tasks (e.g., targets) he will process and which will be handled by the automated device. The problem, of course, is that the user (in most cases) cannot instantly know which of the six situations he or she is in. Making that determination may require some independent problem solving: examining characteristics of the system's performance, assessing one's own capabilities, and/or partly performing the task itself (Lehner, Mullin, and Cohen, 1988). If the user makes poor override decisions, overall performance may be worse than in a system that did not permit override at all. A random override policy in this example yields performance that is better than the user alone, but worse than the aid alone (column 4). Moreover, the process of deciding when and where to override will cost time even if the decisions about whether to override are always made correctly. The result, in a high workload or time critical environment, will either be reduced accuracy by the user when he does override or a delayed response. If we assume a 50% decrement in the user's accuracy in the task when he overrides, an ideal allocation policy produces an overall percent correct of only 57.5% (column 5). This is significantly short of the 61.7% achievable by the aid alone. (Of course, because of the time penalty, the allocation policy would no longer be ideal; the user should in fact hardly ever override.) Finally, a not unlikely possibility is poor override decisions *and* serious decrements in user task performance due to the time consumed by the override decisions. This leads to the worst overall performance: 45% correct (column 6).

On the surface, these considerations support complete automation of tasks whenever the system on average outperforms the user. The task of monitoring a system's performance and occasionally overriding it can be virtually as demanding as performing the task (Curry, 1985; Wiener, 1985). But more considered lessons are also possible:

1. In conditions of low workload or little time stress, the user may be effective at discriminating situations in which he or the aid has an advantage. User control of task allocation between the user and an ATR may make sense, for example, when there are only one or two high priority targets, at relatively long ranges.
2. Under conditions of high workload and high time stress, it may still make sense to allocate some tasks to the user. But the most effective way to accomplish this is for the computer to take the primary responsibility for allocation of tasks. This requires that the computer be well calibrated with regard to estimates of its own accuracy.

Experimental data in an Army air defense identification-friend-or-foe context (Cohen, et al., 1987) support these conclusions. In this experiment, subjects sometimes had access to cues not programmed in the automated algorithm. The ability to override automated conclusions improved performance when workload (i.e., the number of targets) was low, but not when workload was high. Nevertheless, when workload was high, performance could be improved by a *guided override* capability. In this condition, the computer monitored its own confidence in the identification and prompted the user only when certainty was low.

The two modes of interaction identified by Wiener (computer performance with human monitoring, and human performance with computer monitoring) need to be supplemented by a third: computer task performance with *computer* monitoring. In this mode, monitoring the task and intervening in the task are distinguished. The computer monitors itself and alerts the user when it determines that the user might make a significant contribution. The goal is to achieve a more nearly optimal allocation of tasks without the severe performance penalty incurred when the user is responsible for both monitoring the aid and intervening.

2.4.2 User and Computer Mutual Monitoring

The example in Table 1 assumes, for convenience, that tasks are discretely allocated between an automated aid and a user. But it is possible to blend user-computer performance in a more fine-grained and dynamic way. User and computer may perform different components or subtasks of the same task. They may even perform the same components or subtasks and fuse or reconcile the outputs. The distinction in the last section between computer performance and human performance may be better described as a distinction in *balance of initiative*.

Personalization, channeling, and prompting may be regarded as techniques to blend user and computer expertise within subtasks of the same task. But they work somewhat differently depending on whether the user or the aid has the balance of initiative.

Figure 2 illustrates the flow of information and control in a Personalized and Prescriptive interface. The process is personalized in two primary ways: First, adaptive user strategies determine what information the system *displays*, e.g., level of aggregation, generality versus specificity of conclusions, amount of supporting data, alternative conclusions (as discussed in Section 2.2 above). Secondly, the user provides whatever *inputs* (or modifications of system outputs) that he chooses, to whatever degree of completeness or precision. The system fills in whatever gaps are left over, under the constraints of the user's inputs.

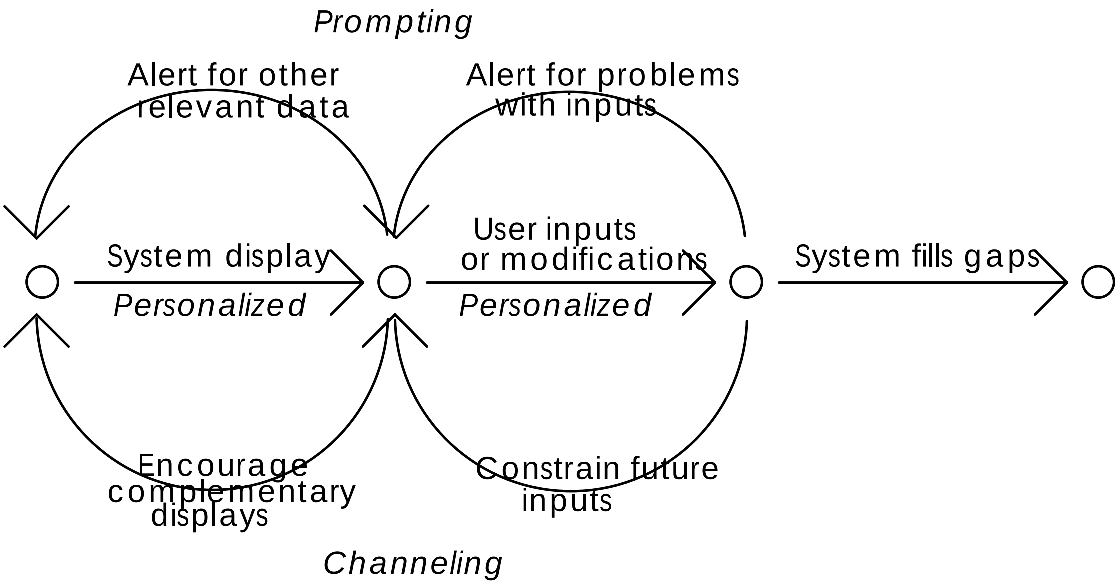
Prompting
Alert for other relevant data
Alert for problems with inputs
System display
Personalized
User inputs or modifications
Personalized
System fills gaps
Encourage complementary displays
Constrain future inputs
Channeling

Figure 2. Personalization, Channeling, and Advisory Prompting during display and user input phases.

Channeling and prompting shape each of these two personalized functions to produce more optimal outcomes, providing non-obtrusive ways for the system's model to influence the user's thought processes. For the display function, channeling encourages the selection of information that typically complements information selected by the user. For the user input function, channeling helps the user keep track of the implications of inputs he has provided, of system conclusions he has accepted, and any constraints these place on future inputs.

Prompting also applies to both personalized functions. In the display process, it alerts the user if information that he has not requested has significant implications for the present problem. (One occasion for prompting would be when other operators with whom the user must communicate are viewing different information.) In the input process, it alerts the user regarding potential problems (e.g., inconsistencies, tradeoffs) in the conclusions he has adopted.

The advisory function itself can be personalized as well as prescriptive: (a) As noted above, it does not require users to abandon their natural modes of problem-solving. Rather, it recommends actions that resemble or mesh with adaptive user-preferred procedures. For example, some ATR users, under conditions of low workload and time stress, may prefer to view relatively unprocessed sensor data rather than system conclusions regarding target classification. The system monitors the information they view and infers the target classification that would be arrived at *based on that data*. If other data exist, not viewed by the operator, which has different implications for classification, the user is prompted regarding that data. Unless he asks for it, however, the operator is not shown the system conclusion. Thus, users are permitted to continue working at the level of concrete data that they prefer. The system functions as a sort of peripheral vision, notifying them of important factors (at the same level of concreteness) that they may have missed. (b) Advisory prompts only occur when the difference between a user-preferred strategy and the solution regarded as optimal by the system is large enough to matter; if differences do not matter, the user is left alone to work the problem his own way. The user himself may determine the frequency with which he receives advice, by determining the size of the discrepancy that would set off a prescriptive prompt. (c) Finally, it may be up to the user whether to accept the system's advice.

The basic structure of Figure 2 is the same regardless of whether the computer or the user has the balance of initiative. But there can be a large difference in the degree of personalization, and the nature of the channeling and prompting. Under conditions of low workload, high stakes, or non-routine tasks, a balance of initiative in favor of the human may be appropriate. The user solves the problem in the way he prefers; so the degree of personalization in both system display and user inputs is high. The primary function of channeling and advisory prompting is to back the user up by guarding against potential errors. Thus, what the computer should be monitoring when the user is in control is the user's decision strategy and the potential errors associated with it.

Under high workload conditions and with relatively familiar and routine tasks, the most appropriate allocation mode may involve a balance of initiative in favor of the computer. In this case, the degree of personalization of both displays and user inputs is less. The amount and type of information displayed will be determined more by the task demands and by system confidence in its own outputs (the system is likely to be better calibrated regarding its own accuracy in relatively routine tasks). Channeling uses prior system conclusions (e.g., that the object of interest is a tracked vehicle) to constrain user inputs. The user's inputs fill in the gaps left by the system (for example, the user determines what type of tracked vehicle the object is). The advisory function in this case requires the computer to monitor its own problem-solving activity, to assess shortcomings (e.g., incomplete data, conflicting cues), and to prompt the user when his contributions are likely to be significant. Thus, what the user should be monitoring when the system is in control is, at least in part, the system's evaluation of its own performance.

Within this framework, human versus computer initiative, and the relative contribution of computer or human to a particular subtask, can be a matter of degree and may shift flexibly as circumstances dictate. Our hypothesis is that displays which take into account adaptive user processing strategies should be more readily utilized, should be understood more quickly and accurately, should be more conducive for eliciting on-the-spot user knowledge, and should lead to overall more effective system performance.

3.0 COGNITIVE PROCESSES IN TARGET RECOGNITION

The starting point for the design of ATR-human interfaces is the set of natural processes and skills that people utilize in recognition. Automated devices that are incompatible with such processes (or that overlook their potential weaknesses) may fail to realize the full potential of user-system synergy. In this section, we examine a series of topics from research in cognitive psychology that seem to have important implications for how human-ATR interfaces should be designed. In each section, after discussing the relevant cognitive research, we try to extract the relevant implications. These implications are stated in the form of research hypotheses. In many cases they are supported by our interviews with attack or scout helicopter pilots or by research in other domains. But in all cases, they should be regarded as preliminary and tentative, subject to more rigorous empirical investigation.

The theoretical framework to be utilized in this section draws upon and integrates two research topics in cognitive science: (i) recognition skills that utilize flexible, structured knowledge representations, and (ii) the role of self-monitoring and self-regulation in strategies that facilitate and verify recognition, control the allocation of attention as a function of workload and goals, and make understanding possible in novel situations.

In Section 3.1 we look at implications of a cyclical, iterative model of visual recognition for the display of visual (and non-visual) data by ATR devices. In Section 3.2, we examine research on verbal categorization for its implications regarding the display of ATR conclusions. In Section 3.3. we turn to research on mental models and explore its implications for the display of uncertain ATR conclusions. Section 3.4 looks at strategies for verifying the success of recognition and their implications for prompting users about ATR uncertainty.

3.1 Cycles of Recognition and the Display of Data

Recognition, in the most general sense, occurs when diverse stimuli elicit a consistent response (Neisser. 1967). In the special case of tactical battlefield ATR systems, recognition involves the categorization of images or image components as tracked or wheeled vehicles, tanks or armored personnel carriers or trucks, T-72s or T-81s, friend or foe, etc.

At least in the near-term, users of ATR systems will need access to sensor information in order to validate or supplement ATR conclusions. Current models of human visual recognition may be relevant to the design of interfaces that provide such information. We will focus on three key features of recognition performance:

- The use of flexible knowledge representations reflecting correlational structure in the domain and internal structure in the stimulus.
- The iterative, dynamic character of recognition, combining top-down and bottom-up processes.

- The utilization of qualitatively different types of information, even within visual processing.

In all three respects, current models of perceptual recognition contrast dramatically with traditional models. For example, according to template-matching models of recognition, an object is recognized by finding a match between the sensory input and a set of prestored representations. Problems with this approach are well-known: (1) An enormous number of different templates would have to be stored in long-term memory in order to cover the natural variability among stimuli falling within the same category. Template-based systems tend to be non-robust: If there is no template corresponding to a novel stimulus, no categorization can be provided. (2) Processing is rigidly bottom-up, with no role for the active direction of attention in order to iteratively resample the stimulus. Both of these problems reflect the fact that template theories fail to represent structure: either the spatial relations within a stimulus or the structure of similarities and dissimilarities among stimuli.

Defining-attribute or feature-matching models improve in some respects on template models. According to defining attribute models, recognition occurs when an object possesses a set of criterial features. Feature models reduce the required number of templates to a small number of features, which reflect dimensions of similarity and dissimilarity among stimuli. The similarity structure posited by such theories, however, is more reflective of artificial laboratory stimuli than the real world. First, in natural categories, recognition can occur even though no single feature or set of features is common to all instances of the category; different combinations of features may be sufficient for the same categorization response. Secondly, real-life features vary in weight; the presence of a particular feature may be a strong category indicator, while many other positive indicators would be required in its absence. Thirdly, membership in the category itself is not all-or-none, but appears to be graded; exemplars differ in their perceived typicality with respect to a category. Finally, spatial relationships among features may be essential for appropriate categorization, but these are neglected.

Feature theories emphasize a bottom-up flow of information from stimulus to feature vector to category. This is a natural consequence of the structural assumptions of this framework. First, since a single, unique set of features is both necessary and sufficient for category membership, there is no rationale for top-down processes that sample *portions* of the stimulus information and redirect attention if the initial sample proves inconclusive. Second, there is no representation of the spatial relations among the features in the stimulus to guide such iterative sampling.

Figure 3 represents a more recent view of recognition, based on Neisser, 1976. It stresses flexible knowledge structures called schemas and the cyclical nature of the recognition process. Like templates, schemas represent prototypical objects or events to which the current stimulus is compared. Like feature models, the schema for a category contains attributes which are characteristic of that category. The attributes, however, may have different weights or importance, and the schema may also specify characteristic relationships among the attributes. Membership in the category is determined by overall similarity to the prototype, not by a specific set of necessary and sufficient features.

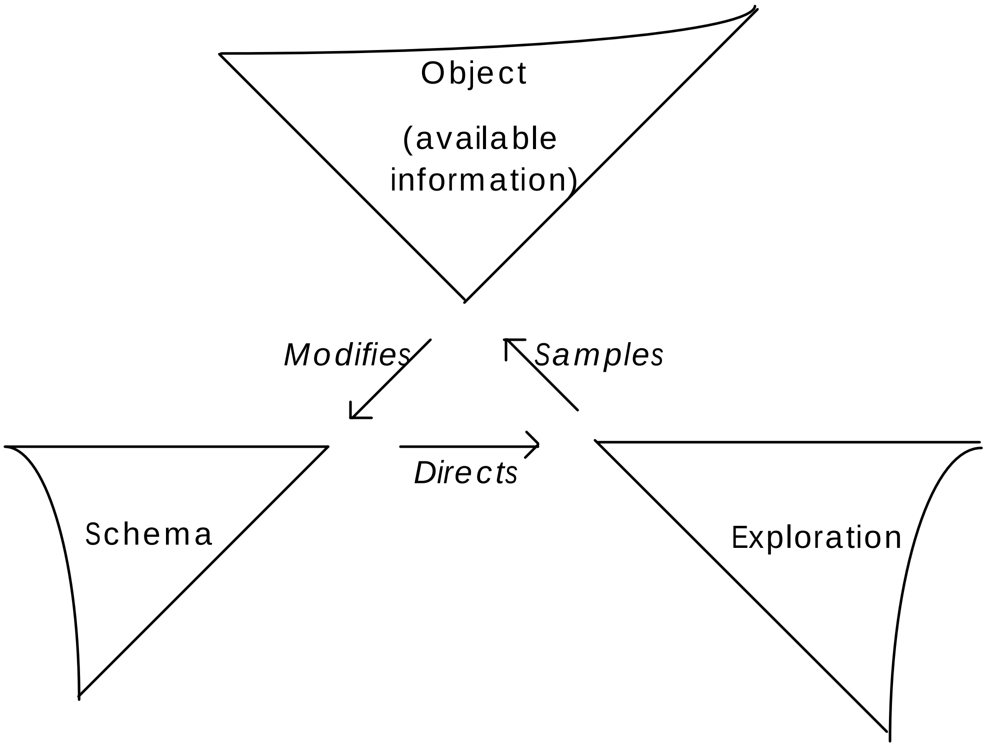

Figure 3. The perceptual cycle (from Neisser, 1976).

Schemas combine bottom-up and top-down processing. When an object matches a sufficient subset of the schema's attributes, the schema is activated, leading to expectations regarding the as yet unobserved features; the schema may thus direct attention to these features or activate exploratory procedures to test for their presence.

A somewhat more detailed view of this cycle reveals the qualitatively different types of information and processing strategies that might be exploited in schema-based recognition. While it would be premature to declare any particular picture definitive, Figure 4, adapted from Kosslyn and Koenig (1992), is based on both neurophysiological and behavioral evidence. These authors summarize evidence for the existence of separate structures and pathways that underlie the modules in Figure 4. First, a stimulus-based attention shifting mechanism draws one's eyes to large changes in the visual field. A visual buffer preserves stimulus information in a spatial structure corresponding roughly to retinal activation. An attention window scans the visual buffer and sends information to two visual pathways for further processing. These two pathways have been referred to as the *what* and the *where* systems, respectively. The first deals with distinctive object properties, which are used both to organize the input into regions and to identify objects. A separate subsystem within this pathway utilizes motion both to segment the input and to identify objects through their distinctive movements. The *what* pathway involves the activation of stored patterns (or visual schemas) that recognize objects by matching the properties and motions of the input.

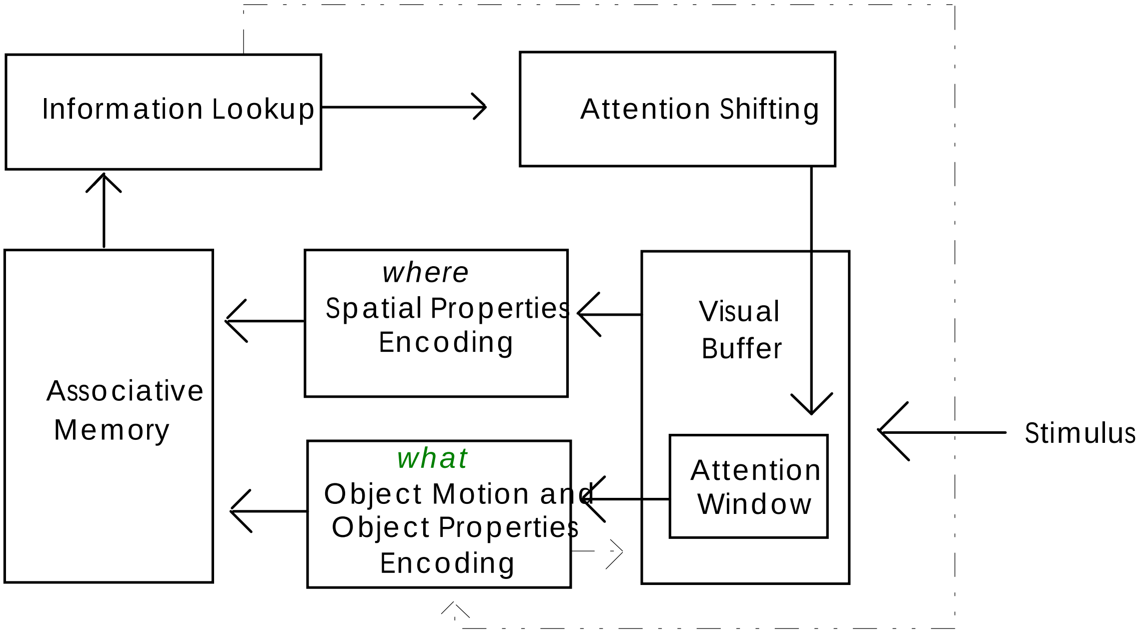

Figure 4. Subsystems used in high-level vision (adapted from Kosslyn and Koenig, 1992).

The second major pathway (concerned with *where* rather than *what*) encodes spatial relations among parts of the stimulus. One subsystem in this pathway computes metric (or continuous) spatial relations between objects or their parts for the purpose of guiding movement. Another subsystem encodes categorical (or discrete) spatial relations between objects or their parts (e.g., the head is at the *top* of the body). The categorical subsystem provides a flexible basis for recognizing objects whose parts may be positioned or oriented in novel ways. The categorical relations among parts or objects may remain constant while the precise details or their metric relations vary.

Associative memory combines information from the object properties system and the spatial system, encoding the relative locations of patterns (properties and movements) with respect to one another and the observer. This permits recognition of objects based on spatial relationships among features, or between the target object and other objects. Higher levels of associative memory also encode features from other sensory modalities, as well as expectations based on prior knowledge. A schema (e.g., corresponding to tank) which best matches this range of inputs may be adopted as a perceptual recognition response.

The information look-up subsystem plays a key role in top-down processing. Based on a preliminary recognition response in associative memory, it utilizes metric and categorical spatial relations to predict additional properties or relationships for the hypothesized object type. It uses these predictions to direct shifts in attention to locations where the properties or relations are expected to be observed.

A final aspect of Figure 4 concerns visual imagery. There is considerable evidence (Kosslyn and Koenig, 1992; Kosslyn, 1980) that imagery utilizes some of the same mechanisms as visual perception. Images may be projected onto the visual buffer, scanned by the attention window, and processed by the same encoding subsystems that analyze perceptual inputs. There is also evidence that imagery may be used in some cases to support perceptual recognition. When highly degraded inputs prevent immediate recognition, an image of a candidate object may be activated on the visual buffer, then transformed in various ways to look for a possible match (Lowe, 1987). The dotted lines in Figure 4 represent the top-down processes by means of which the information look-up system (utilizing information in associative memory) directs the activation of patterns in the visual buffer.

This framework emphasizes the qualitatively different types of information that may be utilized to achieve object recognition under different conditions. More elaborate strategies are called upon as simpler processes fail to achieve a match between inputs and schemas:

- *Basic properties:* In some cases, elementary features of the visual stimulus (e.g., edges, surfaces, and textures) may be sufficient for immediate recognition. The overall shape of an object may be perceived prior to more detailed features (Navon, 1977). For example, a boat-like shape indicates an armored personnel carrier, a truck-like shape indicates a truck; the profile of a low turret and long gun tube may uniquely identify an object as a T62 tank, while the profile of a large turret and medium gun tube identify it as an M60 tank; a triangular pattern of hot spots may indicate a ZIL truck; and so on.
- *Parts and spatial relations:* In other cases, a more detailed perceptual analysis of the object, in terms of its parts and their spatial relations, may be required for recognition. For example, a hot spot may be identified as the hood as opposed to the rear exhaust, by observing that its relationship to other features locates it at the front of the vehicle; the number of wheels may discriminate an M60 from a T62; the slackness of its tracks may distinguish a PT7 from other tanks; etc. Spatial relations with respect to other objects may also contribute to recognition, e.g., the standard pattern of deployment of surface-to-air missile launchers and radars, or the assemblage of trucks, buildings, and antennas at an HQ.
- *Iterative looking:* If the total input from an initial look does not determine a strong recognition response, an operator may look again. In this second look, the operator will be guided by the expected spatial relationships among the parts of known objects. For example, suspecting an M-35 truck, he may look for a high, hot exhaust pipe; if it is not found, he may switch to the hypothesis that the object is a ZIL, which has no such exhaust.
- *Mental images:* For cases of highly distorted or unusual visual inputs, the operator may use visual imagery to generate and transform templates that can be matched to the stimulus. For example, he may imagine how a ZIL would look from an unfamiliar aspect by mentally rotating it; he may imagine the appearance of an M60 with its turret oriented in an unfamiliar way; he may imagine the appearance of one M60 partially occluding another M60; and so on.

- *Non-visual information:* In cases where visual input is highly degraded, very brief, incomplete, or otherwise inconclusive, information from long term memory or other senses may determine recognition in higher-level associative memory. For example, an operator may utilize ELINT data or prior intelligence regarding the presence of tanks rather than trucks to conclude that a visually degraded target is a tank. In some cases, extraneous information of this sort may influence what is "seen" in a top-down fashion (e.g., Palmer, 1975). An operator may actually perceive a very brief, highly degraded, or incomplete image of a truck as a tank because of prior expectations (Bliss, 1974); the presence of a typical surface-to-air missile deployment pattern may cause him to see the characteristic features of a SAM in an object that is in fact something else; the presence of some visual features associated with a ZIL may cause the operator to fill in others that are not in fact visible; and so on.

Implications for ATR interface design: Display of imagery data. A standard display design for target recognition systems includes a large-view imagery display (e.g., based on FLIR or optical sensors) plus small video windows or "chips" that contain larger scale (zoomed in) images of selected objects of interest. The models of human visual processing considered above suggest some preliminary concepts for how data should be displayed within these small video windows by automated target recognition systems. The precise nature of the user-ATR interaction, however, will depend on the balance of initiative between user and automated system.

Computer initiative. Under conditions of relatively high workload or time stress, and in relatively routine tasks (where the ATR is expected to be well-calibrated regarding its estimation of its own accuracy), the primary initiative will be with the computer (as discussed in Section 2.4 above). In this case, the most likely user strategy is to evaluate and possibly override the computer's conclusions. It makes sense, then that the size, number, and nature of the image chips be designed to communicate the basis for an ATR recognition conclusion in the simplest way possible. Such displays, which correspond to the way the perceptual system segments and processes data, will allow operators to quickly and effectively validate the ATR's recommended classification. This process of validation will interfere minimally with the operator's ability to "see the battlefield" as a whole and to perform his other tactical tasks.

Hypothesis 1: If confident ATR recognition is possible based on elementary features of the stimulus (e.g., overall shape, a distinctive pattern of hot spots), the default display size for the video chip can be relatively small. If confident ATR recognition requires analysis of the stimulus into parts, the default size of the chip should be larger. (Other aspects of the video chip display might also be optimized either for perception of elementary features or analysis of parts: e.g., contrast, smoothing, filtering, and edge enhancement.)

Hypothesis 2: If recognition depends on a characteristic movement (or lack of movement) of the object, the default chip should repeat a brief video sequence showing the movement (or lack of movement).

Hypothesis 3: The default chip should include the spatial context of the object of interest if and only if confident ATR recognition depends on that context, i.e., relations of the target to other objects.

Hypothesis 4: Multiple chips for the same object over time should be displayed by default if and only if they reveal crucial features not visible in a single chip.

Hypothesis 5: A template or prototype of the relevant object type should be displayed only if recognition of the object requires the assumption of some transformation: e.g., that the object is viewed from an unusual aspect, that its parts are unusually oriented, that one object is partially occluded by another, and so on. The displayed template should be transformed to produce the closest match.

Hypothesis 6: Non-visual information should be displayed by default only if visual information is insufficient for recognition of the object.

The above hypotheses concern default displays in high workload situations where the ATR is reasonably confident of its own classification. Hypotheses 1 through 6 represent channeling devices whereby the interface directs the user's attention to the most critical features of the stimulus. They provide the most highly diagnostic information to the user regarding the classification. These devices are also personalized in their conformity to human perceptual processes. (When the ATR is not confident in its conclusion, a quite different approach to data displays may be appropriate; these will be discussed below in Section 3.4.)

Even when the ATR is confident in its conclusions, however, users may not always share that confidence. Some users on some occasions may wish to explore the justification for a classification more deeply. This is in effect a form of iterative looking, sampling additional aspects of the stimulus when an initial sampling is unsatisfactory. Thus, an additional element of personalization may be desirable. The user may have the option of requesting alternative types of video displays for a given target. The menu of alternatives from which users choose should reflect the same natural perceptual processes. Options available to operators should correspond to the phases of visual recognition:

Hypothesis 7: Operators should have the option of adjusting chip size at two levels, to optimize for recognition of basic features of the object or for parts of an object.

Hypotheses 8: Operators should be able to select the spatial context of the object, to include related objects.

Hypothesis 9: Operators should be able to select the temporal context of the object to verify the presence or absence of motion.

Hypothesis 10: Operators should be able to request multiple chips for the same target at different points in time.

Hypothesis 11: Operators should have the option of displaying a template of a relevant object type in order to compare it to the image. The template should be capable of user-specified transformations regarding its aspect, occlusion by other object-types, and orientation of its parts.

Hypothesis 12: Operators should have the option of requesting the display of non-image information regarding a target, e.g., ELINT data or prior intelligence.

Human initiative: Under conditions of high stakes and low workload, e.g., with a small number of potential targets at a sufficiently long range, or in non-routine tasks, the primary initiative may be with the user. In this situation the relative priority of personalization and prescription will be reversed. The user will employ his own problem-solving style, with the system serving as an advisor or critic:

Hypothesis 13: Rather than the system determining default displays (as in the computer initiative mode of interaction), the user should specify the types of displays he wants either in general or for a specific target. The choices available to the user should be based on the natural phases of perceptual recognition, as indicated in hypotheses 7 through 12.

Hypothesis 14: The system should indicate by channeling which aspects of the target seem most critical in its own recognition model. For example, in a menu of display options (reflecting hypotheses 7 through 12), the recommended viewing mode based on the system's model could be highlighted.

Hypothesis 15: The system should prompt the user if the classification of the target that would result from the user's preferred viewing mode is significantly different from the classification arrived at by the system's model.

3.2 Basic Categorization Levels and the Display of Conclusions

The outcome of the perceptual processes described in Section 3.1 is a categorization response, e.g., a label that describes the object being recognized. Recent research findings on how people verbally categorize objects may be relevant to the way in which automated recognition systems can most effectively display their own conclusions to human users.

The purpose of categorization is prediction and control. If an object can be categorized based on a sampling of its features, predictions can be made regarding additional features of the object and appropriate responses. Predictions might also be made regarding the features of other objects of the same type. The usefulness of categorization thus depends on objects clustering in terms of shared features (Anderson, 1990) -- or, on correlational structure among features across objects. Categories tend to reflect this correlational structure, grouping together objects that share many features and distinguishing objects that share few features (Rosch et al., 1976).

If prediction is the purpose of categorization, then people should be more likely to use categories that lead to better predictions. Often categories can be arranged hierarchically: e.g., vehicle, tank, T62. Rosch et al. (1976) proposed that in everyday hierarchies (such as furniture, chair, armchair; animal, dog, collie) there is often an intermediate level, called the *basic level*, at which prediction is most effective. Suppose people are asked to list the features that they associate with various categories. Objects that belong to the same basic-level category (e.g., chairs; dogs) share many common features. Most of these properties disappear, however, at the more abstract levels in such a hierarchy (e.g., furniture; animal), where exemplars have far fewer common properties. On the other hand, moving to a more specific level does not produce a comparable increase in features. Most of the properties that collies share are also shared with other dogs.

A comparable phenomenon appears to occur in hierarchies of battlefield classifications. Knowing that something is a tank tells an operator much more than knowing that it is a vehicle; but knowing that it is a T62 adds relatively little additional information. Tank is thus likely to be a basic-level category in this hierarchy. (Exceptions will be noted below; for example, where the task of discriminating friend from foe depends on a more detailed classification.)

Informativeness is, of course, maximized with the most specific categories, but there is a huge increase in the sheer number of categories required to achieve rather small gains in the number of features. Rosch and her colleagues account for basic-level categories in terms of a balance between informativeness (many common features) and efficiency (using as few different categories as possible). Corter and Gluck (1992) formalize these ideas in terms of a measure that trades off the predictability of features from the category (informativeness) against the predictability of the category from features (efficiency).

Experimental data (Rosch et al., 1976, and others) confirm that people are more likely to use basic-level categories in spontaneous naming of objects. They are also faster in verifying that an object belongs to a basic category than to a more abstract or more specific category. (For example, verifying that a robin is a *bird* is faster than verifying that it is a *robin* or an *animal.*) Basic-level category names are the earliest to be acquired by children learning language (Anglin, 1983). Finally, experiments with artificially created objects and category labels confirm that the correlational structure of features influences the labels that subjects tend to learn and use (Corter and Gluck, 1992).

Other observations suggest that the notion of a single, fixed basic level is oversimplified. Rosch et al. (1976) themselves had noted that the basic level depends on expertise; increasing familiarity with a domain (i.e., knowledge of more features) can shift the basic level to more specific categories. It is also likely that tasks requiring distinctions at a more refined level can shift the basic level to more specific categories (Cruse, 1977).

Joliceur, Gluck, and Kosslyn (1984) raised a more fundamental problem. They found that verification that an object belongs to a basic-level category is fast only if the object is a *typical* member of that category. Unrepresentative exemplars are more quickly identified by a more specific category name. For example, a robin is a typical bird, and is quickly verified to be a *bird.* But an ostrich is not a typical bird, and it is faster to verify that an ostrich is an *ostrich* than that an ostrich is a *bird*. Instead of a single basic level, each object may have its own favored level of categorization.

This result forces a revision of the basic-level concept, but it supports the fundamental idea that categories are selected for efficient prediction. An atypical member of a category (such as an ostrich) does not share as many features with other members of the category (e.g., birds) as do typical members (robins, sparrows). For that reason, categorizing an ostrich as a *bird* is less useful than categorizing a robin as a *bird*. It does not as confidently support inferences from some of the features of the ostrich to other features of the ostrich. And it does not as confidently support predictions from features of the ostrich to features of other birds.

Familiarity and goals, like atypicality, can prompt the simultaneous use of different levels of categorization for members of the same class. We may classify people on a baseball field as players, coaches, and umpires, while classifying people in the stands as men or women. We identify our own acquaintances and colleagues (whose behavior we try to understand and predict in detail) by their individual names.

Some authors have stressed the importance of shape for basic-level categories. Barsalou (1991) argues (a) that shape is more rapidly extracted during visual processing than other information, and (b) that most members of a basic-level category (e.g., birds) have the same shape. Members of higher level categories (e.g., animals) have many diverse shapes, while more specific categories (e.g., sparrow, robin) are associated with more detailed visual information than simply shape. This account links basic-level concepts to fixed features of visual processing. But it does not accommodate the apparent role of goals and familiarity in determining the favored level of categorization. It does not explain why categorization may stop at the basic level (i.e., a shape-based categorization) in some cases, but proceed to a more specific level in others. More recently, Barsalou (1992) has suggested a compromise view: Initial categorization may be determined by shape, while subsequent categorization is determined by informativeness and efficiency.

In what follows, the *basic* level of categorization refers to the level in a hierarchy which is generally (but not necessarily always) preferred across all the objects in a domain. The *favored* level of categorization refers to the level that is preferred for a particular object at a particular time. The favored level may not be the same as the basic level if the object is atypical, highly familiar, or part of a task that requires more detailed prediction.

Implications for Display of ATR Conclusions. The existence of basic-level concepts, or favored levels of categorization, can be empirically explored in the domain of battlefield target recognition. Experienced personnel can be asked to list features associated with objects at different levels: e.g., vehicles, structures, aircraft; tracked vehicles, wheeled vehicles; tanks, armored personnel carriers, trucks, automobiles/jeeps; T62s, T72s, etc. A significant increase in the number of features at a particular level, with no comparable increases at more specific levels, suggests the existence of a basic level. Atypical or highly familiar objects may show a relatively large increase in features at more specific levels while typical or less familiar objects do not. To confirm the existence of basic or favored categories for different objects, pictures of those objects can be shown and subjects asked either to name the objects or to verify category labels. Spontaneous naming should involve the favored category label, and verification times should be fastest for the favored category label.

Displaying ATR conclusions at the favored level may facilitate human-ATR interaction. The favored level for a given object corresponds to the category label that is most effective in predicting features of that object; it is also the label that people tend to generate in their own classifications. Thus, providing ATR conclusions at the favored level has both prescriptive and personalized characteristics. The relative importance of personalization and prescription will depend on the balance of initiative between user and automated system.

Computer initiative. Under conditions of high workload and time stress, the ATR will select the default level at which to report its classification conclusions. The chosen level should maximally facilitate user verification of the ATR conclusion by reference to displayed data.

User verification of ATR conclusions might take place in two ways: (a) The user first looks at the data, independently classifies the target, and then compares his classification conclusion with the ATR's. Or (b) the user first looks at the ATR conclusion, generates an image of what the data should look like if the ATR conclusion is correct, and then compares his image with the data. In either case, display of ATR conclusions at the favored category level should facilitate performance: (a) If the user looks at the data first to generate his own classification, his categorization is likely to be at the favored level. Comparison with the ATR conclusion will be quicker if the ATR conclusion is also at the basic level. (b) If the user looks at the ATR conclusion first, then generates an image, the image he generates is likely to be a *typical* exemplar of the category label used by the ATR. If the ATR conclusion is too general, there is no single typical image corresponding to the class, and the strategy of matching an image to the data is likely to result in a mismatch even if the classification is correct. If the ATR conclusion is too specific, generation of a typical image and comparison with the data will be more effortful than necessary.

Hypothesis 16: ATR conclusions should in general be reported at the basic level rather than at more general or more specific levels. We expect that the basic level of categorization will correspond to tank and armored personnel carrier, in contrast to vehicle (too general) or T62 (too specific).

Hypothesis 17: When the target object is atypical or unrepresentative of the basic-level category, the favored level for recognition will be more specific.

Hypothesis 18: If an individual object or a specific object class is highly familiar, or very frequent in a particular battlefield, the favored level for recognition will be more specific.

Hypothesis 19: ATR conclusions should also be reported at the level required for performance of the task. If the goals of the current mission require discrimination at a more specific level, recognition conclusions should be reported at that level as well as at the basic level. If the mission requires discrimination at a more *generic* level than the basic level, conclusions should be reported at the generic level as well as at the basic level.

Hypothesis 20: In some cases, direct recognition at the favored level may not be possible, but recognition at a more specific level is. If part of an object is occluded, or if the object is at an unusual orientation or aspect, normal shape recognition processes may fail. The object may be identified by a more detailed analysis of its parts and their features or relationships (as discussed in Section 3.1 above). For example, most trucks might be identifiable as trucks by their common shape. But if the overall shape of a particular object happens to be obscured, identification of a hot, high exhaust pipe may permit its identification as an M35 (hence, indirectly, as a truck). If recognition of an object requires a more detailed analysis, then conclusions should be reported at the level corresponding to the features that were used. To facilitate verification by the user, the object in the above example should be reported as an M35 rather than as a truck.

These hypotheses concern default displays that channel user attention to the appropriate level of analysis of imagery data. They are also personalized in that they correspond wherever possible to user-preferred levels of description. As in Section 3.1, we have assumed that the ATR is reasonably confident in its conclusions at the recommended level of classification. We discuss issues pertaining to uncertainty in the next section.

Users may differ in their familiarity with targets or in their judgments of typicality. Moreover, as noted in the previous section, on some occasions they may wish to explore the basis for an ATR conclusion more deeply. Thus, an additional element of personalization may be desirable:

Hypothesis 21: Users should be able to request conclusions at a more detailed level than the default display. Such requests should automatically be accompanied by changes in the data display (see previous section), so that information supporting the more detailed classification level can be observed.

Similarly, when users request alternative, more detailed data displays, as discussed in hypotheses 7 through 11, classification conclusions should automatically be reported at the level of detail supported by the requested display.

Human initiative. Under conditions of low workload and time stress, the roles of personalization and prescription will be reversed:

Hypothesis 22: Rather than the system determining the default level at which conclusions are reported, the user should specify the level of categorization he prefers.

Hypothesis 23: The system should indicate by channeling which level of categorization is favored according to its own recognition model. For example, in a menu of options (high level, intermediate level, detailed level), the favored level might be highlighted.

Hypothesis 24: The system should prompt the user if the level of categorization he selects is insufficiently detailed to accomplish the current mission.

3.3 Mental Models and the Display of Uncertain Conclusions

In the battlefield environment uncertainty is inevitable. The available sensor and intelligence data may not unambiguously determine a categorization response at the desired level of specificity. Standard approaches to uncertainty could be utilized in these cases. Such approaches provide displays of alternative possible classifications together with numerical probabilities: e.g.,

Tank - 67%
APC - 20%
Truck - 13%.

There is abundant evidence, however, that such displays may be neither desired by, nor optimal for, human users of ATR systems. We interviewed six active-duty helicopter pilots at Fort Bragg, North Carolina, regarding features of a potential ATR-human interface. Of four pilots who commented on the desirability of displaying alternative possible classifications, only 1 pilot responded positively. Of five pilots who commented on the desirability of displaying a numerical confidence level for the most likely classification conclusion, only 1 responded positively. (The positive answers were given by different pilots.)

There may be good reasons for the pilots' lack of enthusiasm for traditional uncertainty displays. Notice, first, that tank, APC, and truck may each by itself be a basic-level concept, i.e., each of them represents a set of objects characterized by a correlated set of features. But the weighted average of a tank, an APC, and a truck is certainly not a basic-level "category." It is not consistently correlated with perceptual features or actions. A pilot cannot visualize, anticipate the actions of, or prepare for a 67% tank, 20% APC, 13% truck.

Additional insight may be provided by recent cognitive research on the use of *mental models* in reasoning (Johnson-Laird, 1983; Johnson-Laird and Byrne, 1991). According to Johnson-Laird (l983), what distinguishes a mental model from other representations is a close structural isomorphism between the model and the state of affairs it represents. Each object in the model represents an object or set of objects in the world, and relations in the model are analogous to relations among the real objects. Numerous facts can be combined with one another in the same mental model representation without the need for the separate, explicit statement of each fact that is characteristic of logical or probabilistic models. The implications of the combined facts can then simply be recognized or read off the representation without the need for logical deduction or manipulation of formal devices that play no symbolic role.

Perhaps the simplest example of a mental model is a map, which uses symbols and spatial relations to represent objects and their spatial relations. Suppose a pilot knows that howitzer A is west of tank B, and tank B is west of howitzer C, but needs to know the relation between A and C. In a rule-based system, he must represent these beliefs by separate propositions, and then apply a general rule stating the transitivity of "to the west of." But in a mental model representation, such as a map, he places A, B, and C in a single model that satisfies both known constraints:

A ----- B ----- C

The relationship between A and C can now be directly recognized. The relation "to the left of" in the mental model automatically preserves the transitivity of the real-world relation "to the west of."

The relations represented by mental models need not be spatial, but may (for example) be temporal, causal, or conceptual; the relations used in the representation may be, but need not be, the same type as the relations being represented (although they must be isomorphic).

Mental model reasoning is often embodied in mental images. We saw in Section 3.1 above that image-based reasoning may play a role in target recognition, by generating prototype images and transforming them until they match the sensor data. The results of target recognition may then help in the construction of more comprehensive mental models of the battlefield. For example, knowing the types of objects on the battlefield is a pre-requisite for visualizing the degree of threat from their sensors and weapons. Image-based mental models might be used by a pilot in a variety of tasks: i.e., estimating areas where he is vulnerable to anti-air threats and areas where is out of range or masked by terrain, planning a route to avoid the anti-air threats, planning locations and tactics for masking, unmasking, and remasking, estimating locations where he has a good chance of hitting the target, and actually navigating the planned route and executing the engagement. Mental model reasoning may thus play a significant role throughout tactical battlefield situation understanding.

Unfortunately, mental models run into difficulty in the case of uncertainty. A single model cannot display a direct mapping to the existing state of affairs when we do not know what that state of affairs is. Experimental data suggest that problem difficulty increases with the number of separate mental models or possible states of affairs that must be represented (as opposed to the number of steps that would be required by a logical deduction; e.g., Byrne, 1992). Generation, retention, and evaluation of multiple models quickly exceeds human capabilities.

Suppose, for example, that the pilot knows only that howitzer C is east of howitzer A, and that tank B is also east of howitzer A. This cannot be represented in a single mental model. There are two possible models, which vary in the relation between B and C:

A ---- B ---- C

A ---- C ---- B

The pilot must keep *both* of these possibilities in mind if he is not to mistakenly infer that C is east of B (or that B is east of C). If there are a large number of indeterminacies of this sort, a full mental model representation becomes combinatorially intractable.

The requirement of isomorphism between model and reality can be relaxed in various ways to accommodate uncertainty (cf., Johnson-Laird, 1983). Each approach has advantages and disadvantages:

(1) Simultaneously display elements from more than one mental model; e.g., represent uncertainty about the locations of military units by multiple symbols for the same object in the different possible locations:

A ---- (C) ---- B ---- (C)

Similarly, we can represent uncertainty about unit identity by multiple symbols corresponding to alternative possible classifications (e.g., *tank or howitzer or truck*). This design option compromises the natural isomorphism between the spatial representation and a possible real situation. By treating different uncertainties independently, it also fails to capture interactions, e.g., the implications for tactics if object X is a truck and object Y is a howitzer, or if object X is a howitzer and object Y is a truck, if both are howitzers, or if both are trucks. It may be effective when only a very few distinct alternative possibilities are important for the current decision.

(2) Display a single less precise mental model; e.g., represent uncertainty about the locations of military units by spheres or ellipses:

A ---- (B C)

Similarly, we can represent uncertainty about unit identity by a more general category label (e.g., vehicle) indicating what is known with confidence about the unit. Depending on the decision context, even more highly aggregated representations could be used: for example, the display of lethality contours instead of individual threats. The same lethality might represent any of a large number of combinations of locations and capabilities of enemy forces. Representations of this kind are not at the favored level of categorization; they do not communicate the information that is typically needed for decisions about avoiding threats, unmasking, and engaging targets. They do not give people a feeling of confidence regarding their understanding of the situation. But they may be of value when the current decision or action does not in fact depend on more precise information.

(3) Assume a single model and revise it later if necessary; e.g., select a single threat location (or identity) from among those that are possible:

A ---- C ---- B

This preserves the isomorphism between the representation and a possible situation at a favored level of representation. Indeed, human reasoning does often seem to be assumption-based: A single model is tentatively adopted (often, the worst case), subject to revision in case of new data (Cohen, 1989; Johnson-Laird, 1983; Doyle, 1979; Reiter, 1980). Dynamic adjustment in response to feedback and new information replaces exhaustive up-front analysis (Connolly and Wagner, 1988). The danger, of course, is that the decision maker may lose track of the assumptions that have been made and feel an unwarranted sense of certainty regarding the represented possibility (e.g., Einhorn, 1980).

Among the assumption-based strategies that people seem to adopt are: assuming the worst case, assuming that a familiar sensor or source is reliable until proven otherwise, assuming that a new source is unreliable until proven otherwise, and assuming that a piece of data means what it usually means.

A rather large literature confirms that people do not integrate all possibilities the way standard normative methods dictate. For example, research by Cohen, Leddo, and Tolcott (1988) found that attack submarine commanders estimated ambiguous quantities (such as the range to a target or the probability of being counterdetected by the target) by selecting one concrete possibility rather than by combining diverse possibilities into a weighted average. For example, they assumed that the worst-case (i.e., the closest) range estimate was correct and all others wrong, and they assumed that the target had the best possible sensor capability rather than an "average" capability, when deciding how long they could safely approach an enemy submarine. A worst-case planning strategy may make sense: It guarantees that the selected option will produce an adequate outcome no matter what the true situation. Unfortunately, it can also lead to overlooking important opportunities: options that are only slightly worse in the worst case but significantly better if the worst case is false.

Assumption-based reasoning can also lead to overconfidence. People often seem to suppress awareness of alternative possibilities once they are able to construct a single mental model of how an event could occur. For example, people consistently overestimate the probability of expected events, even in areas where they are experts (Kadane and Lichtenstein, l982).

Implications for the display of uncertain conclusions. ATR displays must not only support target recognition; they must also be compatible with the other key decisions (e.g., route planning, navigation, targeting and engagement) that pilots make in order to carry out their mission. Explicit displays of alternative recognition conclusions (especially with numerical probabilities) may detract from the pilot's ability to construct effective, transparent mental models of the battlefield within which he must act. Perhaps for these reasons, the pilots we interviewed were almost unanimously opposed to such explicit uncertainty displays.

Computer Initiative. Under conditions where the computer takes the initiative, three design questions must be addressed: (1) How to display uncertain ATR recognition conclusions? (2) When and how to prompt regarding the existence of uncertainty itself? (3) How much image and non-image data should be displayed? We address the first of these issues in this section, and the other two in the section that follows.

In Section 3.2 we recommended that ATR conclusions be reported down to at least the basic level, and that conclusions should be even more specific (i.e., at the favored level) if demanded by the atypicality of the object, its familiarity, the requirements of the mission, or the need for unusually detailed analysis of the data. But what if the ATR is significantly uncertain at the level of specificity that would normally be appropriate?

Recent research suggests that the accuracy of an ATR system influences its perceived value to pilots (Becker, Hayes, and Gorman, 1990). In one experiment, pilots saw little value in a system with a correct detection rate below .7 or a false alarm rate below .167 per square degree. In our own interviews, we found large individual differences in the confidence thresholds that pilots thought appropriate. Preferred thresholds for the display of recognition conclusions ranged from 51% to 95%. Some pilots wanted to see ATR conclusions no matter how uncertain (thus setting a threshold at 51%), while others preferred not to be bothered with any conclusions that were not virtually certain (thus setting the threshold at 95%). There was also variability as a function of the scenario and mission. A lower threshold for reporting conclusions (e.g., 60%) might be appropriate for deep attack missions, where no friendlies would be expected. A higher threshold (e.g., 90%) might be required for a close-in battle, where friendlies and hostiles are intermixed.

What then should be *the* threshold for displaying ATR conclusions? We argue that there is no single answer. The most effective criterion for displaying recognition conclusions will depend on the situation. In particular, it depends on: (1) the expected cost of a recognition error (e.g., the likelihood and cost of shooting a friendly) and (2) the time available for making a decision.

In the hypotheses that follow, we postulate an ATR interface with a minimal "intelligent" component that is capable of gauging these two variables in a fairly coarse and qualitative manner. For example, *high stakes* situations include discriminating friends from foes in close battles, and discriminating significant threats (e.g., anti-air artillery) from non-threats. An example of a *low stakes* recognition problem involves discriminating different types of targets from one another when all are acceptable for engagement. Available time might be assessed by the ATR as a simple function of distance to threats and threat weapon ranges. The following hypotheses are summarized in Table 2 below.

Hypothesis 25: When stakes are low, classification conclusions should be personalized according to the preferences of the pilot, regardless of the available time. As noted above, some pilots prefer ATR conclusions only down to the level of specificity at which the ATR is confident. For these pilots, the ATR will display a generic category (e.g., vehicle) when further specification (e.g., tank versus truck) doesn't matter - for example, because the vehicle is too far away to be a threat, or in a location where the mission precludes engagement. Such a display preserves the isomorphism of a mental model representation without providing a spurious sense of certainty. Other pilots prefer displays down to the favored level (e.g., tank), even if the ATR is relatively uncertain. These pilots may be more effective framing a mental model at a consistent level of description that matches their knowledge.

Hypothesis 26: In high stakes situations, i.e., where information at the favored level *is* needed for mission tasks, the appropriate display depends on the available time. If time stress is not severe, then the ATR should display classification conclusions only down to the level of specificity at which it is confident. This enables the pilot to provide an independent judgment regarding the object's classification. Prior exposure to the ATR's uncertain conclusion may influence the way the pilot visualizes the object, and result in less accurate recognition performance.

Hypothesis 27: Suppose that information at the favored level is required for mission tasks, but time stress is severe. The pilot is engaged in other tasks, and/or the target is a threat at close range. In this case, the ATR should display a single conclusion at the level of specificity required by the task. The displayed conclusion is arrived at in part by data and in part by assumption. For example, a plausible strategy is to display the worst-case recognition result as long as it is reasonably likely given the data. Such a display enables the user to construct a simple mental model of the situation, and also guarantees a minimum outcome of his decisions. In this situation, the requirement for timely action outweighs the advantages of independent recognition judgment by the pilot.

These displays channel the user's attention under conditions of moderate to high stakes. Under low time pressure, they encourage him to provide an independent judgment. Under high time pressure, they encourage him to adopt the most appropriate recognitional assumption. Table 3 summarizes the recommendations regarding display of conclusions.

	Time Stress Low	**Time Stress High**
Stakes Low	Personalized display of conclusions - generic or specific (Hyp. 25)	
Stakes Moderate	Generic conclusion (Hyp. 26)	Favored-level conclusion (Hyp. 27)
Stakes High	Generic conclusion (Hyp. 26)	Favored-level conclusion (Hyp. 27)

Table 3. Display of uncertain conclusions.

Additional elements of personalization, however, can increase the synergy between human and ATR:

Hypothesis 28: The user should have the option of displaying the alternative possible conclusions at whatever level of specificity he desires. This capability permits pilots to check their own independent conclusions against the conclusion of the ATR, when the initial ATR display was generic (Hypothesis 26), or to explore alternative possible assumptions when the initial ATR display involves an assumption (Hypothesis 27). In designing an in-flight situation assessment aid for the Air Force, we found that pilots preferred a display that assumed the worst-case when there was conflicting evidence regarding the identity or location of a threat, but also wished to have the option of viewing best-case or most-likely-case displays (Cohen, Tolcott, and McIntyre, 1987).

In the introduction to this section, we listed three ways in which mental model displays could be modified to accommodate uncertainty: (1) simultaneous display of alternatives, (2) display of less precise models, and (3) tentative assumption of a single precise model. Hypotheses 25 through 28 spell out different conditions under which each of these approaches might be appropriate.

3.4 Verification Strategies and Uncertainty Prompts

Reasoning with mental models, as we observed in the last section, has both strengths and weaknesses. When generic mental models are used to represent uncertainty, they may fail to provide necessary information; and precise mental models that are based on assumptions may prove to be misleading or false. To guard against these potential pitfalls, mental models may be combined with strategies for critiquing or verifying the adequacy of the model. ATR-human interfaces can be designed to support such verification strategies.

A growing body of research on expert problem solving may provide some insight into the nature and role of these verification strategies. The initial emphasis in this field was on general purpose problem-solving strategies, such as breaking a complex problem down into simpler components or working backwards from goals to means (e.g., Newell and Simon, 1972). In the past two decades, however, interest has focused on skills that are domain-specific and knowledge-intensive rather than general-purpose and analytical. Empirical studies comparing novices and experts in fields such as physics, chess, and computer programming have supported a view of expertise as the accumulation of direct responses to familiar situations (e.g., Chase and Simon, 1973; Larkin et al., 1980). Experienced problem-solvers "recognize" the problem and "see" what to do. Expert recognition is effective because of the character of the knowledge upon which it is based. Recent research has emphasized the structure of expert knowledge rather than simply its quantity: for example, many studies suggest that recognition by experts occurs in terms of fundamental domain concepts rather than superficial features of a problem (Chi et al., 1981; Shoenfeld and Herrman, 1982; Weiser and Shertz, 1983; Adelson, 1984; Larkin, 1981). This is consistent with our earlier discussion of object recognition (Section 3.2) in which we emphasized the use of categories that capture the correlational structure of the environment.

There is also evidence that experts are more skilled than novices in strategies that control how recognition is achieved and validated. (Larkin, 1981; Glaser, 1989; Brown and DeLoache, 1978). These strategies increase the flexibility of expert performance and enhance its ability to deal with novel situations, where data are incomplete or conflicting. Thus, while experts may be said to "recognize" familiar problems, the process of recognition itself is by no means simple or invariant, but is subject to a variety of choices based on intermediate results of the process.

One type of strategy increases the chance of successful recognition. Experts change their representation of the problem until it makes contact with their knowledge, i.e., until it becomes "familiar." According to Larkin et al. (1980), a sketch of the superficial objects and relations in a physics problem is often constructed and examined in order to determine the next step: If the depicted system is familiar, the expert may proceed directly to the equations required for solution. If the system is unfamiliar, the expert constructs an idealized representation (i.e., a free-body diagram), which is then used in the generation of solution equations. In extremely difficult problems, moreover, experts appear to shift to a more novice-like strategy of means-ends analysis (Larkin, 1977). These strategies are consistent with our discussion of the use of imagery in object recognition (Section 3.1): When an object is not immediately recognized, experts may generate and transform templates (e.g., corresponding to possible articulations of the object's parts, or occlusion by other objects) until a match is achieved. Just as problem solving is now being seen as a kind of recognition, object recognition can be seen, for some purposes, as a form of problem solving.

Strategies also play a role *after* the problem has been recognized and (apparently) "solved." Physics experts were found by Chi et al. (1982) to utilize the abstract physical representation of the problem to verify the correctness of their method and result, e.g., by checking whether all forces are balanced, whether all entities in the diagram are related to givens in the problem, etc. Experts are highly practiced in deciding how much and what kind of checking is required. Similarly, in medical diagnosis, Patel and Groen (1991) found that experts generated accurate diagnostic hypotheses early, "and spent the rest of the time evaluating in order to confirm and refine the diagnosis...," while novices had difficulty evaluating their hypotheses. In mathematics, novices, unlike experts, often fail to verify their results against prior knowledge (Riley, Greeno, and Heller, 1983). These strategies extend our discussion of the iterative processes in visual perception (Section 3.1): Additional kinds of information (e.g., relationships among parts, relationships to other objects, non-visual data) may be consulted even if recognition has been achieved based on earlier samplings, to the extent that the initial recognition is not conclusive.

Under time constraints, experts attempt to consult the most important information first. For example, in an Army air defense identification-friend-or foe context, Cohen, Leddo, and Tolcott (1988) found that as target density increased, experienced operators shifted to strategies that emphasize breadth over depth. They examined fewer classification cues per target, while continuing to examine all contacts. Such a strategy assures at least some high-value information about each potential threat.

Experts and novices also appear to differ in the way they verify information from external sources such as advisors or computer-based aids. Cohen (1992) found dramatic differences between experienced and inexperienced commercial airline pilots in how they handled dispatch advice in the face of weather uncertainty. These findings may have implications for the way Army pilots should respond to uncertain ATR conclusions.

In Cohen's (1992) experiment, active-duty commercial airline pilots had to make a now-or-never decision to divert or not to divert to a third airport in the face of an uncertain weather condition moving in on the original destination and alternate. The potential severity and degree of uncertainty of the weather were varied. The company dispatch advised the pilots either to divert or to continue (dispatch advice was a between-subjects variable, and was orthogonal with respect to the seriousness of the weather). Dispatch advice had no effect at all on the decisions of pilots who had less than 20 years of commercial flying experience. They looked at the weather information and made a decision, either adopting a worst-case strategy or trading off potential risks against potential gains. By contrast, dispatch advice had a significant effect on the decisions of pilots who had more than 20 years commercial flying experience. Many of these pilots appeared to take dispatch advice as a starting point and look for potential problems. If none were found, they followed the advice. This strategy allowed them to efficiently allocate their attention to relevant information (For example, if dispatch recommended continuing, they looked only at worst case predictions regarding the destination and alternate. If dispatch recommended diversion, they looked at best case predictions and if those were good, at worst case predictions). The strategy also allowed them to take advantage of dispatch advice without following it blindly.

A similar verification strategy might be effective in Army pilot interactions with an ATR. Particularly under high workload conditions and relatively routine tasks (where the ATR is accurate and well-calibrated regarding its own accuracy), the most efficient strategy may be to take the ATR conclusions as a starting point and look for problems. In that case, the ATR's assessment of confidence in its own conclusions assumes great importance. Uncertainty serves as an alert that directs the pilot's attention to the recognition task *per se* and to the recognition of particular targets, causing him to sample data efficiently in a search for potential problems.

Implications for the Display of Uncertainty Prompts. There are two primary reasons for explicitly alerting the user of an ATR to the existence of uncertainty:

1. To affect decisions directly. For example, when a target is identified as a foe but could be a friend, recognition confidence is important in deciding whether to engage immediately or to spend more time and accept more risk collecting additional data. When a target is identified as a non-threat but could be a threat, recognition confidence is important in deciding whether to continue with a route that puts the aircraft within range of that target or take more time to avoid it. Avoidance of *false alarms* and *misses* is thus a crucial consideration in the design of ATR-human interactions.

2. To orient the user's attention to a recognition task where he may be more successful than the ATR. As discussed above, pilots might use the ATR as a starting point, and attempt to verify or critique results that the ATR regards as uncertain.

Traditional uncertainty displays, which include all alternative conclusions together with numerical probabilities, are not warranted by either of these needs. With respect to (1) discouraging immediate engagement or routing decisions, traditional displays add little or nothing to the information that the classification of a hostile or a threat is uncertain. With respect to (2) orienting the pilot's attention, they merely delay the pilot from actually attending to the object that needs to be recognized.

Hypothesis 29: Uncertainty prompts should be binary: A conclusion is either uncertain or it is not. The threshold for display of such a prompt, however, can vary with the situation.

Hypothesis 30: Uncertainty indicators should not interfere with visual examination of the target, since an important function of such prompts is to direct the pilot's attention to the stimulus. Interference may occur due to blinking the target or enclosing the target within a rectangle (which may distort perception of the target in context). Appropriate indicators might include distinctively shaped symbols, blinking symbols, or distinctively colored symbols placed adjacent to the target.

The display of uncertainty prompts reflects a tradeoff between the potential costs and likelihood of recognition errors and the time available for making a decision:

Hypothesis 31: If time stress is low, the ATR should provide an uncertainty prompt for the user whenever the likelihood and the cost of an error are moderate to high.

Hypothesis 32: If time stress is high, uncertainty prompts should be provided only if the likelihood and the cost of a recognition error are high. The alternative possible recognition results must have major implications for the success of the mission.

MacMillan (1992) speculates that a pilot might use one of two possible verification strategies: scanning for possible targets identified by the ATR (checking for false alarms) and scanning for possible targets that have been missed by the ATR (checking for misses). The prompts described above help the user allocate his attention in a reasonable way between these two extremes. Instead of focusing on just false alarms or just misses, he might be on the look-out for a false alarm with regard to one target, and a miss with regard to another. The occurrence of an uncertainty prompt is keyed to the actual likelihood and cost of different kinds of classification errors for each individual target.

Nevertheless, we expect that false alarms will receive more attention than misses in some battlefield situations, and that misses will receive more attention than false alarms in other battlefield situations. In a close-in battle, there is expectation that friends and hostiles will be intermixed. If the ATR incorporates this type of information in its reasoning, the calculated stakes will be highest, and the expected costs of errors greatest, for targets that the ATR classifies as hostile but which could be friends. These kinds of targets will be most likely to get uncertainty prompts, and the interface will therefore guide the user to monitor, in effect, for false alarms. In deep battles, by contrast, few friends are expected. The stakes will be highest, and the costs of errors greatest, for targets that the ATR classifies as non-threats (e.g., trucks or structures) but which could in fact be threats (e.g., tanks, anti-air artillery). These kinds of targets will be most likely to get uncertainty prompts, and the user will be guided, in effect, to monitor for misses.

The following table summarizes the implications of these hypotheses regarding prompts for uncertainty, along with the previous hypotheses regarding display of uncertain conclusions, as a joint function of time stress and the expected cost of a recognition error. Relevant hypotheses are in parentheses.

	Time Stress Low	**Time Stress High**
Stakes Low	Personalized display of conclusions - generic or specific (Hyp. 25)	
Stakes Moderate	Generic conclusion + prompt (Hyp. 26, 31)	Favored-level conclusion (Hyp. 27, 32)
Stakes High	Generic conclusion + prompt (Hyp. 26, 31)	Favored-level conclusion + prompt (Hyp. 27, 32)

Table 3. Display of uncertain conclusions and prompts.

Implications for the Display of data. A final design issue, arising when the ATR is uncertain of its own conclusions, concerns the data that should be displayed to the user. In Section 3.1 we discussed appropriate displays of imagery (e.g., in video windows or chips associated with specific targets) and non-imagery data to support rapid user validation of ATR conclusions. The basic principle in that discussion was that displays should be as simple as possible, following a natural human visual processing sequence, while still being sufficient for validation of the automated recognition conclusion. If the user questioned the system conclusion, he could "look again," by requesting more detailed or more comprehensive displays. When the ATR is uncertain in the first place, however, and if the conclusion matters for the mission at hand, the user's role may go beyond validation. The user is now actively looking for problems with the ATR's conclusion, and the user's own judgments are more likely to significantly enhance the accuracy of the classification. To support this potential user contribution, more extensive data displays will be required. For the user to go beyond the ATR, he must have access to data that is not selectively filtered through the ATR's recognition process.

The purpose of the uncertainty prompts described above was to direct the user's attention to recognition problems where he might do the most good (within the overall context of battlefield goals). The most effective ATR data display will therefore reflect the same tradeoffs between the stakes of the decision and the available time that were relevant for displays of prompts:

Hypothesis 33: In just those situations where the user is prompted regarding uncertainty (i.e., low time stress and moderate to high stakes, or high time stress and high stakes; see Table 3), the ATR should display all data that might have any bearing on the classification of the target. In particular, data of each qualitative type described in Section 3.1 should be displayed if any *signal* is detected in the corresponding data type. For example, a large chip size should be displayed if there is any detectable differentiation among the parts of the object (even if the ATR cannot determine their relevance for the classification). The spatial context should be displayed if any detections occur in the vicinity of the target object (even if the ATR cannot discern their relevance). The temporal context should be displayed if any movement of the target object is detected. Multiple snapshots of the target should be displayed if any change in the target is detected. Any available non-imagery data (e.g., ELINT or prior intelligence) should also be displayed.

Hypothesis 34: In all other situations (i.e., when stakes are low, or when stakes are moderate but time stress is high), the ATR should display only the data types that it has actually utilized to support its conclusion, as described in Section 3.1.

Hypothesis 35: As in Section 3.1, the user should have the option of requesting any additional information of any type.

Under some conditions, the ATR might support a slightly more extensive problem-solving process by the user. At the user's request, it might provide appropriate templates for comparison with the displayed data:

Hypothesis 36: In cases where time stress is low and stakes are high, channeling should make it easy for the user to view alternative target-type templates (e.g., for tank, armored personnel carrier, truck) and compare them for a match to the stimulus. The user should have the option of transforming the templates in order to improve the match to the stimulus. Transformations might include altering the aspect of the templated object, altering the orientation of its parts, or visualizing one object type occluding another.

Table 4 summarizes these hypotheses concerning data display:

	Time Stress Low	**Time Stress High**
Stakes Low	Display only supporting data (Hyp. 34)	
Stakes Moderate	Display all potential signals + template options (Hyp. 33, 36)	Display only supporting data (Hyp. 34)
Stakes High	Display all potential signals + template options (Hyp 33, 36)	Display all potential signals (Hyp. 33)

Table 4. Display of data under ATR uncertainty.

Human initiative: When time is more ample (e.g., because there are very few potential targets and they are at relatively long ranges), human initiative and a higher degree of personalization may be appropriate. Under these conditions pilots would choose the kind of data as well as the types of conclusions that are displayed. Of the six pilots we interviewed, four expressed a preference for ATR displays based on confidence, while two preferred ATR displays based on a particular level of categorization regardless of confidence.

Hypothesis 37: Pilots should be able to set the threshold of confidence required for an ATR conclusion. The ATR will then present conclusions down to the level of specificity permitted by its confidence (but not more specific than the favored level). Generic category labels (e.g., vehicle; target) will be used when the ATR is uncertain.

Hypothesis 38: For these pilots, prompts should be presented when categorization to a more specific level is required for successful performance of the mission.

Hypothesis 39: Alternatively, pilots should be able to set the level of categorization at which the ATR displays its conclusions. The ATR will then present conclusions at that level regardless of uncertainty.

Hypothesis 40: For these pilots, prompts should be presented when there is significant uncertainty at the user-specified level of categorization and when the uncertainty might result in costly errors.

Note that the role of prompting in the two cases is complementary. When the pilot chooses to display conclusions in terms of confidence (Hypothesis 36), prompting occurs with reference to level of categorization. For example, suppose the pilot wants conclusions displayed only if they achieve a 90% confidence level. If only *vehicle* can be displayed at the desired level of confidence, the ATR may prompt that a decision between tank and truck is required for the mission. On the other hand, when the pilot chooses to display conclusions in terms of category level (Hypothesis 38), prompting occurs with reference to confidence. For example, suppose the pilot wants conclusions displayed at the level of *tank.* The ATR may identify an object as a tank, but alert the pilot that the relevant object is known with high confidence only as a vehicle.

4.0 EXPERIMENTAL PLAN

In this section we lay out a preliminary plan for testing the key hypotheses from Section 3.0. In accordance with the Personalized and Prescriptive aiding methodology (Section 2.0), each cycle of research will proceed in two phases. The first phase tests fundamental ideas about human information processing in the context of tactical battlefield target recognition. These ideas, about how humans naturally perceive and classify objects, handle uncertainty, and monitor for false alarms and misses, underlie the hypotheses about human-ATR interface design. Based on the results of these experiments, then, the interface design hypotheses will be revised and/or fine-tuned. (Some of these hypotheses concern adaptation to user strategies and others concern protection against potential pitfalls of those strategies.) In the second phase of each research cycle, the interface concepts will be tested directly, to assess their likely impact on overall user-ATR performance.

In Section 4.1 we describe an approach to performance measurement that can be used both in the study of cognitive processes and in the evaluation of interface design concepts. The measurement procedure captures tradeoffs that threaten to confound assessment of performance quality, both between time and accuracy and between false alarms and misses. At the same time it provides a tool for investigating the qualitative character of user information processing strategies.

Section 4.2 describes the experimental tasks to be performed by subjects. These tasks correspond to the specific areas that we plan to investigate, which were discussed in Sections 3.1 through 3.4. They include: the sequence of stages in visual information processing (Section 4.2.1), favored levels for classifying objects (Section 4.2.2), use of mental models and assumptions in handling uncertainty (Section 4.2.3), and verification strategies that monitor for misses and false alarms (Section 4.2.4).

Section 4.3 presents an overall plan and schedule for research. The research is divided into two parts, corresponding approximately to the first and second years respectively. The first year will focus on the display of data and conclusions (discussed in Section 4.2.1 and 4.2.2), while the second year will focus on displaying uncertainty and ATR prompts to support user monitoring (discussed in Sections 4.2.3 and 4.2.4). Each year will comprise a full cycle of research, beginning with underlying cognitive principles and concluding with ATR interface designs. Finally, Section 4.4 discusses experimental stimuli, the experimental environment, and subjects.

4.1 Measures of Performance

4.1.1 Recognition Accuracy

In target recognition, the observer categorizes a target by assigning it to one of *n* classes (e.g., tank, APR, truck). In simple detection tasks, by contrast, the observer assigns a stimulus to one of two classes: signal (e.g., target) versus noise (no target). In detection, only two kinds of errors are possible: missing a target that is present and reporting a target that is not present. In target recognition, the number of possible confusions among *n* classes totals *n*(*n*-1). Despite these differences, some of the same methodological issues arise in assessing detection and recognition performance. A brief discussion of detection will thus serve as an introduction to the assessment of recognition performance.

Measures of detection performance that focus on only one kind of error (misses or false alarms) may misrepresent the quality of performance, by overlooking changes in the other variable. Observers can achieve higher accuracy with respect to either type of error by accepting a higher rate of error of the other type. Neither measure by itself, therefore, adequately captures the underlying ability of the observer to discriminate between targets (signals) and non-targets (noise).

Signal detection theory (SDT) is a powerful method for dealing with the tradeoff between false alarms and misses in detection performance (Green and Swets, 1966). Signal detection theory has two components: (1) The representation of the internal effects of the stimulus as a random variable on a single underlying dimension. The internal effects of signal and noise are usually assumed to form independent Gaussian distributions, with the mean of the signal distribution offset from the mean of the noise distribution by a fixed amount. The distance between the means of these distributions, called *d'*, is a measure of the observer's ability to discriminate targets from non-targets. (2) The application of decision theory to these internal representations. The internal effect of any particular stimulus has a certain probability of coming from the signal distribution and a certain probability of coming from the noise distribution. The ratio of these two probabilities is the likelihood ratio for that stimulus, expressing its value as an indicator of a signal. Based on the expected overall frequency of signals and noise, and the costs of the two different kinds of possible errors, the rational observer establishes a threshold on the underlying continuum. If the internal effect of a stimulus falls above this threshold, it is reported as a signal; if the internal effect falls below the threshold, it is reported as noise. The criterion for reporting a signal is called β; shifts in β reflect changing tradeoffs between misses and false alarms in different situations.

Estimates of *d'* and β can be derived from a Receiver Operating Characteristic, or ROC curve, which plots hits against false alarms. ROC curves can be generated by several methods: (a) by varying payoffs and costs for hits and false alarms; (b) by varying the prior probabilities of signals and noise; (c) under a single payoff and probability condition, by assuming normal, independent equal-variance distributions; and (d) under a single payoff and probability condition, by asking observers for assessments of confidence in the presence of a signal.

Recognition tasks are far more complex than simple detection tasks. (We assume in this research that a target has been detected. In principal, however, detection and recognition could be regarded as a single, even more complex recognition task; i.e., "no target" could be included as one of the classes to which a stimulus might be assigned.) In recognition, potential tradeoffs between different kinds of confusion errors occur between every pair of classes to which a target may be assigned. For example, an observer can always increase his chance of correctly identifying tanks if he is willing to accept an increase in the chance that he will misidentify APRs and trucks as tanks. At the same time, he can increase correct identifications of trucks relative to APRs by allowing an increase in the number of APRs misclassified as trucks. The observer's *ability* to discriminate can also vary between different classes, independently of his biases: For example, he may be better at discriminating trucks from tanks and APRs than he is at discriminating tanks and APRs from one another.

In principle, signal detection theory could be extended to recognition with multiple categories. In place of a single decision dimension, there must be a space defined by separate axes for each category. Multivariate distributions would correspond to the internal effects of each category within this space, and also to the internal effect of no target. Confusability between categories would be determined by the distance between the means of their corresponding distributions. Prior probabilities of different categories, and the costs of different kinds of confusion errors, might influence the way observers partitioned this space into decision regions corresponding to each category.

The analysis of experimental data in terms of a generalized SDT model is prohibitively complex (e.g., Broadbent, 1971). However, a variety of alternative techniques exist that under many conditions (e.g., assuming normal and symmetrical internal sensory dimensions) produce very close approximations to the SDT analysis. These techniques also accomplish the basic goal of distinguishing the effects of discriminability from the effects of decision criteria in multi-category tasks. Some of these alternative techniques also offer advantages in their interpretability in terms of underlying psychological processes.

The best known technique of this kind is the Biased Choice Model described by Luce (1956, 1977). According to this model, the probability of a response in the presence of a stimulus is a joint function of response bias and similarity to the stimulus associated with the response. More formally, the probability $p(i,j)$ of a response j in the presence of stimulus i is proportional to the bias $b(j)$ in favor of response j, and the similarity $\eta(i,j)$ between classes i and j. Recognition performance can thus be characterized by a matrix:

		Responses				
		1	**2**	**3**	**...**	**n**
Stimuli	**1**	*b(1)*	*b(2)η(1,2)*	*b(3)η(1,3)*		*b(n)η(1,n)*
	2	*b(1)η(1,2)*	*b(2)*	*b(3)η(2,3)*		*b(n)η(2,n)*
	3	*b(1)η(1,3)*	*b(2)η(2,3)*	*b(3)*		*b(n)η(3,n)*
	...					
	n	*b(1)η(1,n)*	*b(2)η(2,n)*	*b(3)η(3,n)*		*b(n)*

It is assumed that similarity is symmetric, that the similarity of a stimulus with itself is 1, and that the sum of the *b(i)* is 1. The following equation describes response probabilities:

$$p(i,j) = \frac{b(j)\eta(i,j)}{\sum_j b(j)\eta(i,j)}$$

Bishop, Fienberg, and Holland (1975) provide a maximum likelihood method for estimating the parameters of this model, and assigning confidence intervals to the parameters, from a confusion matrix describing the number of responses of each type to stimuli of each type.

Another approach to modeling recognition performance also provides a good approximation, under certain conditions, to signal detection results, but is based on a specific process model. The Informed Guessing Model (Pachella, Smith, and Stanovich, 1978; also, related models by Broadbent, 1971, and Townsend, 1971) is a special case of the biased choice model, in the sense that it fits only a subset of the data that might be fit by the less restrictive biased choice model.

According to the Informed Guessing Model, when a stimulus is presented, a number of perceptual events may or may not occur in a probabilistic manner, each of which narrows down the class of possible responses to the stimulus. These events might correspond to detections of relevant features of the stimulus, or use of prior evidence about the stimulus. After the set of possible responses has been narrowed in this way, the observer guesses from the set of surviving responses. Response biases influence these guesses. If no information about the stimulus is extracted, the confusion set (from which the observer guesses) consists of all the possible stimulus categories. If the confusion set is narrowed down to one stimulus class, no guessing is necessary.

The probability of response *j* to stimulus *i* is:

$$p(i,j) = \frac{B(j)}{B(i)+B(j)}\xi(i,j) + B(j)g$$

$$p(i,i) = 1 - \sum_{i \neq j} p(i,j)$$

g is the probability that insufficient information was extracted from the stimulus to rule out any class of stimuli. *ξ(i,j)* is the probability that enough information was extracted only to narrow down the possibilities to *i* or *j*. *B(i)* is the bias in favor of response *i*. Again, similarity is assumed to be symmetrical, i.e., the probability of any particular confusion set is the same regardless of which stimulus actually led to that confusion set. The sum of the bias parameters equals 1. Finally, the sum of the probabilities of all the confusion sets containing any particular stimulus class must equal 1:

$$g + \xi(i) + \sum_{j \neq i} \xi(i,j) = 1$$

Notice that the probability of a correct response can be more directly represented as a sum of "true discriminations" and various kinds of guesses:

$$p(i,i) = \xi(i) + \sum_{j \neq i} \frac{B(i)}{B(i)+B(j)}\xi(i,j) + B(i)g$$

This leads to a simple correction for guessing or bias, in which *p(i,i)* is replaced by *ξ(i)* as a measure of the ability to discriminate i from other stimuli.

Maximum likelihood methods for assessing the parameters of the Informed Guessing Model are described in Pachella, Smith, and Stanovich (1978).

The Biased Choice Model and (in a more explicit way) the Informed Guessing Model permit examination of the similarity structure among a set of stimuli, and investigation of how that structure might change over the time course of visual processing. Moreover, the bias parameter in these models permits examination of assumptions that observers make in the absence of conclusive evidence.

4.1.2 Recognition Speed

We observed in the last section that exclusive focus on false alarms or misses may lead to incorrect conclusions about the quality of performance. A similar tradeoff can occur with respect to speed and accuracy. Observers are able to reduce errors (of either kind) by spending more time to process a stimulus. Conversely, observers can improve their reaction time to a stimulus by accepting a degradation of accuracy. Thus, an exclusive focus on accuracy measures alone (or speed measures alone) may also produce a misleading characterization of performance quality. For example, a manipulation that improves accuracy may do so only by increasing time; if so, no overall improvement in quality can be inferred. Speed is particularly important in battlefield target recognition. A large number of targets, or the urgency of a potential threat, may prevent the pilot from processing stimuli to the maximum level of accuracy. The most significant contribution of ATRs may in fact be in situations where the operator himself does not have adequate time to fully process a stimulus.

A speed-accuracy operating characteristic is a plot of accuracy against the amount of time spent processing. One way to obtain such a characteristic is to impose a penalty that increases with delay, and to impose costs for inaccuracy, and to vary the relative size of the two kinds of penalty. In the battlefield target recognition context, a penalty for delay might correspond to an increasing chance of being fired on as the aircraft remains exposed. Penalties for inaccuracy include the costs of firing on friends or the costs of missing threats. Another way to obtain a speed-accuracy operating characteristic is simply to impose a response time-window (Reed, 1976) or deadline (Pachella and Pew, 1968) on the observer, and to measure accuracy as the window or deadline is varied.

Increasing the time deadline (or decreasing the penalty for delay) should improve accuracy but have no effect on response bias. Thus, in terms of the accuracy models described in the previous section, *d',* η *,* or ξ should change as a function of time, but β*, b,* or *B* should not. This constitutes a check on the construct validity of the accuracy model.

A plot of overall accuracy (e.g., percent correct classifications) against time can help validate an ATR design. If accuracy with an ATR is as good as accuracy without an ATR at all response times, and better at some of the response times, then we can safely conclude that improvements in accuracy occur without increasing the time required for processing the stimulus.

Perhaps the most important contribution of the speed-accuracy methodology recommended here is the insight it will provide into the qualitative time course of perceptual processing. By plotting the specific components of accuracy (e.g., the $\xi(i,j)$ and the $\xi(i))$ against time, we get a series of snap shots of processing, revealing which features of a stimulus set are extracted early and which are extracted late, and how recognition processing might be qualitatively affected by the introduction of an ATR. Such an analysis can be an important and innovative input into the design of the ATR interface.

4.1.3 Engagement Task Performance

Recognition does not take place in a vacuum. It occurs in the service of other decisions, such as selecting a target or avoiding threats, upon which battlefield success depends. Selection of an appropriate target in particular depends on accurate and timely classification. A major concern of this research will be to test hypotheses about improving recognition performance in this larger context: We will ask how displays that improve recognition speed and accuracy can be optimized to improve the chances of successful engagement.

The simplest relationship between classification and engagement decisions occurs when a single target is detected, and the operator must decide whether or not to engage it based on its classification (For example, engage it if it is an enemy tank; otherwise, do not). Two kinds of error are possible: Killing a non-target (a false alarm) and failing to kill a target (a miss), and each is associated with its own cost and probability. In this situation, a straightforward analysis is possible, e.g., within Signal Detection Theory (cf., Section 4.1.1 above). The operator should engage if the likelihood ratio (i.e., the probability of the evidence given it is a target divided by the probability of the evidence given that it is not a target) exceeds a decision criterion. The decision criterion is determined by the costs of the different kinds of errors and the relative prior probabilities of targets and non-targets (cf. MacMillan, 1992). In this situation, the engagement decision simply *is* the classification decision.

A more challenging situation arises when the operator encounters multiple possible targets, but has only one engagement opportunity -- e.g., because of limited weapons or the danger of prolonged exposure. He must therefore select the *best* target for engagement. This case is more difficult for two reasons: (1) Multiple targets afford more opportunities for error. If ATR classification accuracy is 90%, and the subject must prioritize four possible targets, the probability of classifying all 4 targets correctly is only 66%. The probability of classifying 10 targets correctly is 35%. Even when classification accuracy is high for individual targets, the probability of selecting the best target for engagement can be surprisingly low. (2) Each potential target adds another type of potential error. Even if the operator engages an acceptable target, there is the possibility that he missed a greater or more immediate threat, possibly to his own survival.

In this situation, the operator can benefit if he rapidly and accurately integrates information from all the potential targets.

To assess engagement decision performance, we use the following paradigm:

N potential targets are presented on each trial. One of these is to be selected for engagement based on its classification and the associated degree of threat. For example:

- the mission is to engage tanks that pose a threat to friendly forces; there are several tanks at roughly the same distance; but they differ in their weapon range, hence, in the immediacy of the threat they pose. A unique one of these objects is the best target for engagement, because it has the longest weapon range, on any given trial.
- the mission is to engage hostile vehicles; there are several tanks in the vicinity; but only one is hostile.

The observer is given varying periods of time, *T*, to respond, as in the deadline method described in the previous section. One simplification in this paradigm is that the observer cannot choose to remask and then pop up again to collect more information. He must choose a target, or choose not to engage, based on the information collected within the given time.

A simple measure of performance in this context is the probability of selecting the best target for engagement plotted as a function of the available time. (Another possibility would be to leave time unconstrained, and measure the time required to achieve a given level of accuracy.) This engagement paradigm will be used to test interface concepts with respect to display of data (Section 4.2.1), display of conclusions (Section 4.2.2), and uncertainty handling (Section 4.2.3).

The probability of choosing the best target will depend not only on the observer's ability to classify targets, but on the way he goes about allocating his attention in the more complex engagement task. In Section 4.2.4 below, we describe some simple verification strategies, heuristically model their effects on engagement performance, and describe research to explore their implications for ATR displays.

4.2 Experimental Procedures

4.2.1 Sequence of visual processing

The first set of experiments addresses the question of how much and what type of supporting data should be provided for subjects regarding targets classified automatically by the ATR. In Section 3.1 we proposed a set of design concepts based on a model of the processing of visual and non-visual information. According to that model, recognition begins with crude or global features of a stimulus and progresses to an analysis of its component parts and their relations, spatial relations of the object to other objects, visually transformed models of the object, and non-visual information. Hypotheses 1 through 6 draw implications of this model for the display of data about targets.

These hypotheses apply to situations in which the user is relatively confident in ATR conclusions, e.g., when the ATR's declared confidence is high and the ATR is well-calibrated in assessing its own uncertainty. (We consider situations where uncertainty is crucial in Sections 4.2.3 and 4.2.4 below.) For such targets, the interface design should be optimized to facilitate rapid validation of the ATR conclusion. According to hypotheses 1 through 6, the basis for an ATR conclusion should be communicated in the simplest way possible, i.e., by means of information that is extracted early in visual processing -- as long as such information makes the conclusion sufficiently probable. Hypotheses 7 through 12 provide the operator options for exploring additional information if he chooses.

The experimental approach to this topic proceeds in two phases: first, verification in a realistic target recognition context of the basic hypotheses about psychological processes; second, validation of the suggested ATR interface design concepts.

Procedures to assess psychological processes. We propose a quite general and powerful method for analyzing the stages of visual and non-visual processing, based on the speed-accuracy tradeoff tools described in Section 4.1.

Are global target features extracted before detailed parts? To answer this question, we need a stimulus set that is structured with respect to both a global and a detailed feature. For example, consider a simple set of four stimuli, in which only two relevant features are visible - the overall shape of the object and the number of its wheels:

General Type	Specific Type	Global feature	Detailed feature
APC	**BMP**	boat shape	6 wheels
APC	**BTR**	boat shape	4 wheels
Tank	**M60**	tank shape	6 wheels
Tank	**T62**	tank shape	4 wheels

Subjects will be asked to identify the specific type of these stimuli, i.e., respond to each stimulus from the set of categories, *BMP, BTR, M60, and T62*. Moreover, these responses will be subject to a response deadline. For example, in one block of trials, the subjects must respond within 350 milliseconds; in another block of trials, they must respond within 550 milliseconds; and so on. Different types of confusion errors can then be plotted against the amount of time available for processing the stimulus.

If overall shape is extracted before a detailed feature like the number of wheels, there should be more confusion errors between BMP and BTR and between M60 and T62, than between any other pairs of stimuli (e.g., BMP and M60, BMP and T62, etc.). This difference, moreover, should be greatest for brief processing durations, and diminish or disappear for longer processing durations.

The performance measurement models discussed in Section 4.1 permit a rigorous analysis of this effect. For example, using the Informed Guessing Model, we can plot the probabilities of different confusion sets (the $\xi(i,j)$ and the $\xi(i))$ against time. Our prediction corresponds to the hypothetical results in Figure 5 (based on actual data with other kinds of stimuli, reported in Pachella, Smith, and Stanovich, 1978). In these data, confusions between stimuli that differ in overall shape are entirely accounted for by being in the pure guessing state. Thus, at any given time an observer is at one of three levels of processing:

1. the confusion state consisting of all four categories, i.e., he has extracted no relevant information,
2. one of the confusion states BMP+BTR or M60+T62, i.e., he has extracted only overall shape information, or
3. one of the four states consisting of a single category, i.e., he has extracted overall shape information *and* detailed part information.

As the time available for processing increases, the probability of being at level 3 increases and the probabilities for levels 1 and 2 decline.

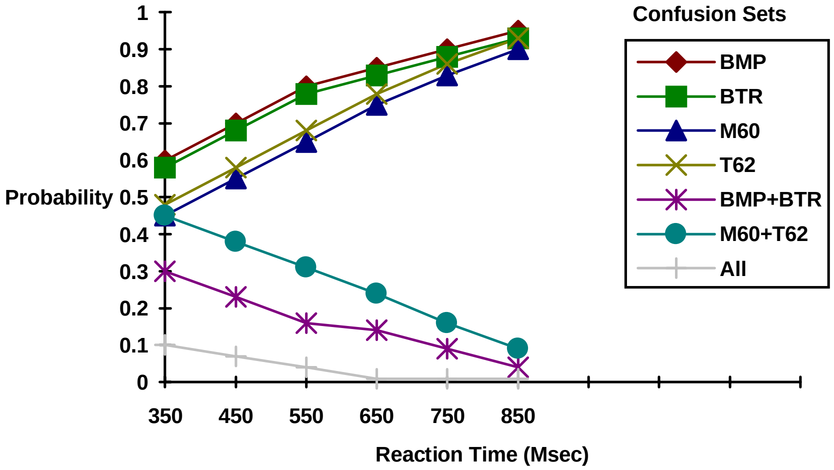

Figure 5. Probability of confusion sets as a function of time.
(Hypothetical data)

Note that in this example, overall shape is correlated with general type: Tanks have a similar shape, as do APCs. On the other hand, there is no type label that corresponds to the detailed feature, *four wheels*, or the detailed feature, *six wheels*. Thus, it might be claimed that the reliance on overall shape earlier in visual processing is due to its association with familiar labels rather than to a characteristic of visual processing as such.

But this objection would not invalidate the importance of the finding that overall shape is processed early. In fact, it is exactly what would be expected based on the discussion in Section 3.2, where we noted that basic level categories (e.g., tank, APC) are often associated with a common shape. The *reason* for the priority of shape may, therefore, be its association with basic-level categories. The causality, however, might run in both directions: One reason for basic categories could be their association with shape. In any case, the *fact* that overall shape is processed early would retain its importance for ATR interface design.

Nevertheless, there are elementary visual features that may be processed early but which are *not* associated with type labels. Thus, the association of early processing with labels can be tested by selecting sets of stimuli that are structured in the appropriate way. For example, the pattern of hot spots in a FLIR image of a T62 (tank) and a BTR (APC) may be similar; the pattern of hot spots for a BRD (APC) and a ZIL (truck) may also be similar. We predict that salient patterns of hot spots will often be processed early even if the resulting confusion sets do not correspond to labeled categories.

Our discussion has focused for illustrative purposes on the priority of overall shape versus detailed parts of an object in recognition processing. The same basic procedure, however, can be used to test the priority of processing other sources of information regarding a target. Sets of targets and response categories will be structured in appropriate ways so that specific discriminations among target categories can only be made by the observer's use of object motion, spatial context, previous sightings of the same object, parts that are articulated or oriented in an unusual way (therefore, presumably requiring internal transformation of a template), and non-visual information. The occurrence of appropriate confusion sets will then indicate, as in the example above, which discriminations have been made early and which have been made late in the course of recognitional processing.

Finally, we will be alert to the possibility of individual differences in the priority of different types of visual and non-visual information. These differences may result from differences in experience; our interviews, for example, suggested that less experienced pilots may tend to focus more on details than on global properties. Differences might also arise from the types of experiences to which different pilots have been exposed: e.g., night versus day flying, short-range versus long-range weapons, or theaters where different target properties assume importance for classification.

Procedures for validating ATR design concepts. The results of the experiments on visual processing will be used to fine-tune and revise our hypotheses regarding ATR interface design. These hypotheses will then be implemented in an interactive computer mock-up of an ATR interface. The display concepts will be directly tested by means of this mock-up.

Whereas the initial experiments were narrowly focused on the perceptual components of target recognition, the interface validation studies will widen the scope and increase the ecological validity of the experimental environment. They will place subjects in a relatively realistic simulated battlefield context, which replicates at least some of the actual tasks and mission priorities of Army helicopter pilots. The relevant performance measure in this context is not only recognition accuracy *per se*, but the way ATR displays support or disrupt engagement prioritization. Moreover, the time constraints imposed in this study are not artificially associated with individual targets, but with overall accomplishment of the engagement prioritization task. Finally, these tests will address not only effects on performance, but also effects on confidence and user preferences.

The tests will involve recognizing and prioritizing targets for engagement. Subjects will be presented with 5 or 6 targets that have been detected and classified by the ATR with a high degree of confidence (e.g., 90%). The subject's task involves validating the ATR classification and selecting one of the targets for engagement based on the classification. The time available to perform the task will be varied, e.g., from 2 to 40 seconds. Mission engagement priorities will involve the type of target (e.g., enemy tanks) or a combination of target type and battlefield location or activity. Target detections and classifications will be indicated by the ATR at the level required by the mission, for example, by a symbol placed adjacent to the target in the large-view imagery display. In a certain percentage of the cases (e.g., 10%), the ATR classifications will be incorrect in a way that is relevant to engagement priorities.

Three conditions will be compared which differ in how video windows or chips are used to provide additional information about the potential targets:

1. No additional information is displayed about the targets.
2. Selected additional information is displayed, in accordance with the ATR interface design hypotheses. In other words, information is displayed in the sequence of natural visual processing, until enough information is available to classify the target.
3. Full information about each target is displayed (e.g., large size chips, spatial context, past sightings that provide additional cues, transformed templates, and non-visual data).

The primary performance measures will be the frequency with which subjects select the appropriate target for engagement as a function of the amount of time available for the task. A secondary measure is user confidence in the selection of targets. Speed-accuracy predictions are illustrated by the hypothetical data in Figure 6.

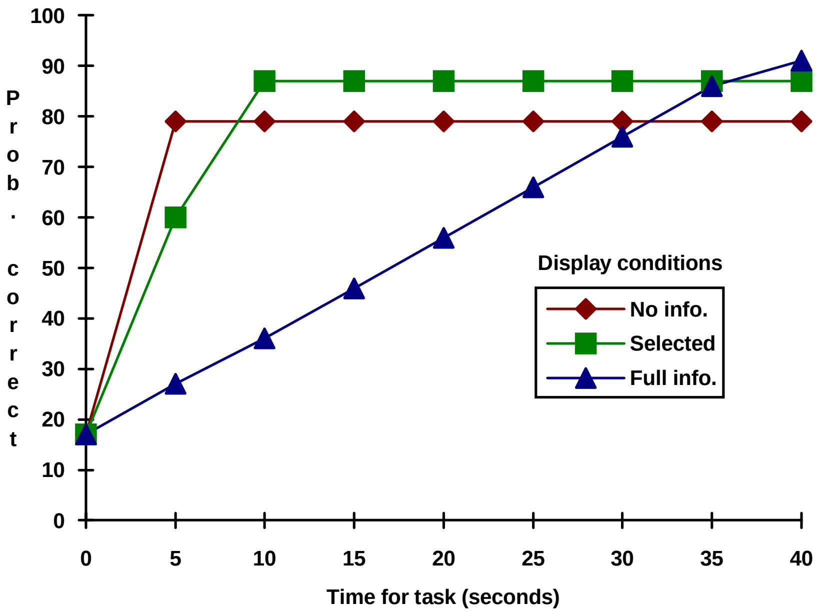

Figure 6. Hypothetical data in different data display conditions.

The most salient fact about Figure 6 is a simple tradeoff: The better a condition does under extreme time stress (i.e., when the time for the task is only 5 seconds), the worse it does when there is more time available. The reason is straightforward: The no-information condition provides nothing to distract the operator in time-stressed conditions, but it also provides nothing to help him, if he does have some time, to confirm or disconfirm the classification conclusions of the ATR. Thus, the maximum accuracy achievable by the no-information condition is limited by the original classification accuracy of the ATR. This is attained rather quickly (we have assumed about 1 second per target for optimal performance.) The full-information condition, by contrast, provides too much information in time-stressed conditions; this results in less efficient selection and utilization of the crucial information (including the basic ATR conclusions) and lower accuracy. On the other hand, given enough time, the full-information condition allows the user to second-guess and possibly improve the accuracy of the ATR classification. Optimal performance, however, takes much longer to achieve (we have assumed about 8 seconds per target).

The selected-information condition is less disruptive than the full-information condition when time is scarce, and is more useful than the no-information condition when time is available. While the no-information condition may outperform it at extremely brief task durations, and the full-information condition may outperform it slightly at unusually long task durations, we expect that the selected-information display may do best at the majority of task durations in between.[1] The argument for utilizing the selected-information display, then, is based on the highly favorable tradeoff it achieves between speed and accuracy.

The number of targets, may be varied to explore the robustness of any conclusions regarding the superiority of selected-information displays. The larger the number of targets, the less information observers can utilize regarding each target. This would be predicted to increase the relative advantage of the no-information condition with respect to the full information condition. (Later, when we turn to our discussion of verification strategies (Section 4.2.4), we will explore another crucial variable, ATR accuracy. The less accurate the ATR, the more advantage there will be in displaying fuller information for selected targets.)

We expect that confidence judgments by pilots will roughly correspond to accuracy. In the no-information condition, we expect confidence to be low, since pilots in our interviews have consistently expressed a desire to take a more detailed look at targets selected for engagement. We also expect confidence to be low in the full-

[1] Figure 6 is based on a specific verification strategy: examining all information about each target before going on to the next target (described in Section 4.2.4 below). In the no-information condition, the maximum success rate for selecting the most appropriate target is 79%, based on ATR classification accuracy of 90% per individual target plus the possibility of guessing when targets are misclassified and more than one target appears eligible for engagement. In the selected-information condition, we made the somewhat arbitrary assumption that ATR errors could be corrected 40% of the time when there was sufficient time to examine all the information; when fed into the model, this resulted in a maximum engagement success rate of 87%. In the full-information condition, we assumed that if all the information were examined, ATR errors could be corrected 60% of the time; this resulted in a maximum success rate for that condition of 92%. Finally, we made assumptions about the rate of examining targets with different amounts of information: 1 target per second for the no-information condition, .6 targets per second for the selected-information condition, and .13 targets per second for the full-information condition. Similar conclusions could be derived from other verification strategies and a fairly wide range of parameter values.

information conditions, when pilots are presented with large quantities of information which they have insufficient time to examine. Finally, we expect preference by pilots among these conditions to follow along the same lines.

Hypotheses 7 through 12 in Section 3.1 deal with pilot options to display more information for specific targets. An additional condition, therefore, might be: Selected information plus pilot options. We predict that affording pilots the opportunity to explore questionable classifications in more detail may result in higher accuracy, especially when time is available. This effect would be more pronounced, also, the less accurate the ATR.

The pilot-options condition offers a more fine-grained measure of pilot preferences. By tracking the actual information requests of pilots, we can validate our hypotheses about the sequence of processes in recognition. We can also look for effects of individual differences as well as the effects of differences in the validity or importance of different types of cues.

4.2.2 Level of categorization

The question to be addressed in this set of experiments concerns the appropriate level of generality or specificity at which ATR conclusions should be reported to users. In Section 3.2 we proposed design concepts based on theories in the area of verbal categorization. According to those theories, there is a basic level in hierarchies of linguistic terms, which people are quickest to learn, are more likely to use in naming objects, and are fastest to verify. Concepts at the basic level tend to capture the correlational structure among features across objects, without unnecessary proliferation of category terms. The basic level may shift as a function of expertise. Moreover, category terms at more specific levels are more likely to be used for objects that are atypical members of the basic-level category, for objects that are highly familiar, and when the goals of the task dictate a more detailed description. We described these effects in terms of a *favored* level of categorization. Hypotheses 16 through 20 describe how displays of ATR conclusions might vary as a function of these variables. Hypothesis 21 provides users with the option to display conclusions at other levels when they chose to do so.

We propose two phases of experiments: first, verification in the target recognition context of hypotheses regarding favored levels categorization; second, validation of the suggested ATR interface design concepts.

Procedures to test psychological model: A number of experimental procedures will be used to verify the role of basic and favored level concepts in target recognition:

(1) *Generation of attributes:* Subjects will be presented with category names (e.g., vehicle, tracked vehicle, tank, T62) and asked to generate as many features of the category as possible. As concepts become more specific, more features should be associated with the category label. The increase in number of features should be relatively great going from a more general level concept to the basic level. But the increase in number of features should be relatively small in going from the basic level

to more specific levels. For example, the following features might be associated with tracked vehicle, tank, and T62, respectively:

Tracked vehicle	**Tank**	**T62**
Tacks	Tracks	Tracks
Wheels inside tracks	Wheels inside tracks	4 wheels inside tracks
	Tank shape	Low profile tank shape
	Turret	Low turret
	Gun	Large gun
	Exhaust hot spot	Exhaust hot spot
	Engine hot spot	Engine hot spot

Knowing that an object is a tank implies a lot of information about that object - far more information than is implied by knowing only that it is a vehicle. The additional information that the object is a T62 does not produce a comparable increase in information. For the most part, the new information consists of refinements and constraints on the features already associated with tank: e.g., while all tanks have guns, a T62 has a relatively large gun. The same pattern recurs in the hierarchy wheeled-truck-KRAZ:

Wheeled vehicle	**Truck**	**KRAZ**
Wheels	Wheels	Wheels
	Truck shape	Truck shape
	Cab	Cab
	Cargo	Large cargo
	Front engine hot spot	Front engine hot spot
	Differential hot spot	Differential hot spot
		Bridging unit bet. cab & cargo

Figure 6 illustrates this effect (We have given .5 credit for attributes that are only refinements of existing features).

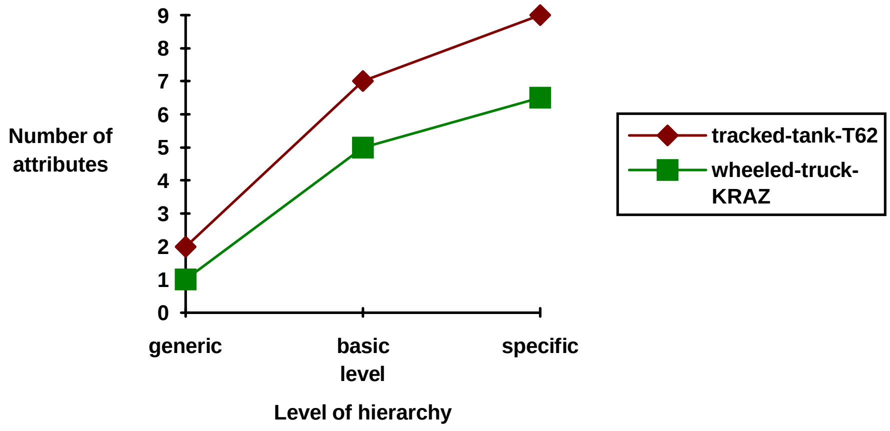

Figure 7

(2) *Judgments of typicality.* A category and a list of instances of the category will be presented (e.g., tank - T62, M60, PT7; truck - KRAZ, M35, ZIL; etc.). Subjects will be asked to judge the typicality of the instances as members of the category on a scale from 1 to 100. We predict that instances will tend to be judged non-typical when they represent a larger than average increase in the number of associated features in procedure (1) above.

(3) *Spontaneous naming.* Subjects will be presented with images of objects and asked to classify them. They will be permitted to select a classification response from any desired level of generality (e.g., wheeled, truck, KRAZ). With no further constraints on the task (e.g., in terms of mission), we predict that the most commonly used labels will be from the basic level, as defined in (1) above. The more atypical the object, the more likely subjects will use more specific labels, as in (2) above.

(4) *Verification times.* Subjects will be given images of objects together with ATR classifications of those objects and will be asked to verify the ATR classification. We predict that verification times will be fastest when classifications are at the basic level, as defined in (1), or - for atypical objects - when classifications are at a more specific level, as defined in (2).

We will be alert to the possibility of individual differences in the basic level, as a function of different amounts of experience as well as differences in the tasks and missions making up that experience. Such differences should be reflected consistently by findings in procedures 1, 2, 3, and 4.

Procedures for validating ATR interface concepts. The results of the experiments on verbal classification will be used to fine-tune and revise the hypotheses regarding ATR interface design. These hypotheses will be implemented in the computerized ATR mock-up and directly tested.

The basic experimental context will be the same as described in Section 4.2.1 above. The tests will involve a realistic battlefield task: recognizing and prioritizing multiple targets for engagement. Subjects will be presented with 5 or 6 targets that have been detected and classified by the ATR with a high degree of confidence (e.g., 90%). The subject's task involves validating the ATR classification and selecting a single target for engagement based on the classification. The amount of time available for the task will be varied, e.g., from 2 to 40 seconds. A fixed amount of visual and non-visual data about each target will be displayed.

We propose to manipulate the following variables:

- the mission, and thus the level of categorization that is relevant for engagement decisions,
- the level of categorization of the ATR conclusion,
- the frequency of occurrence and typicality of category instances,
- the level of detail of the data that must be consulted in order to classify the object, and the associated level of categorization.

Let us initially consider only the first two variables (i.e., the favored level will be the same as the basic level of categorization). The mission-related categorization level will be either generic, basic-level, or specific. The level of categorization of the ATR conclusion may also be generic, basic-level, or specific. In addition, the ATR conclusion can be reported *both* at the basic level *and* at the mission-required level (if they are different). Thus, we have the following combinations of these two variables:

Mission		**ATR Conclusions**		
Generic	Generic	Basic-level	Specific	Generic + Basic*
Basic-level	--	Basic-level*	Specific	--
Specific	--	--	Specific	Basic + Specific*

(In this study, the ATR conclusion will always be at least as specific as the mission level: There is little point in testing conditions where the ATR is capable of confidently classifying a target at the mission-required level, but chooses not to.)

The asterisks represent predicted best performance: i.e., the level of ATR conclusion that we expect to be optimal for the indicated level of mission (Hypotheses 16 and 19). A common element of these three predictions is display of conclusions at the basic level. We expect that basic-level displays, regardless of the actual mission requirement, will facilitate the observer's validation of ATR conclusions. If he looks at the data first, spontaneous generation of a category label for comparison with the ATR conclusion tends to involve the basic level. If he looks at the label first, production of a mental image for comparison with the data is also most effective at the basic level.

If the mission-required level and the basic level differ, however, an additional step is required. When the mission requires classification at a more generic level than the basic level, the next step is simply verbal: e.g., recalling that a tank is a vehicle. When the mission is more specific than the basic level, the next step may involve additional processing of the data, e.g., to discover what kind of tank is present. Display of the mission-level categorization by the ATR in addition to the basic level supports this subsequent step in processing.

We now consider the third and fourth variables listed above. These can cause the favored level of categorization to be different from the basic level. If a specific type of tank occurred far more often than any other in a particular battlefield setting, the specific label (e.g., T62) might become the favored level for that type of tank. We predict that performance will tend to be better with specific labels for highly frequent instances (Hypothesis 18). Similarly, a highly atypical instance of a basic category may be more appropriately labeled at a specific level (Hypothesis 17).

In some cases, as pointed out by Hypothesis 20, recognition of the object cannot take place at the basic level. Because of a degraded image, occlusion, an unusual aspect or orientation, etc. processing to a more detailed level of classification must take place first, and identification at higher levels must be inferred from the more specific category. We propose to examine this situation under two levels of mission:

Mission	**ATR Conclusions**		
Basic-level	Basic-level	Specific	Basic + Specific*
Specific	--	Specific*	Basic + Specific

Asterisks represent our predictions regarding the best level to display ATR conclusions for the given level of mission. If the mission requires classification at a specific level, *and the data also demand classification at a specific level,* there is no point in displaying the basic level conclusion. Spontaneous naming in response to these atypical data will tend to be at a more specific level rather than at the basic level; and imagery in response to a basic-level label would not be useful in verifying the ATR's conclusion with these data.

According to Hypothesis 21, users might have the option of requesting conclusions at more detailed levels. Another test might involve providing users this option, in conjunction with the automatic display of the corresponding, more detailed visual and non-visual data (Hypotheses 7 through 12). We predict that this capability will improve the operator's ability to detect and correct ATR errors, especially under conditions where ample time is available.

4.2.3 Mental Models and Uncertainty

The questions to be addressed by this set of experiments concern how ATR conclusions should be displayed when the ATR is uncertain. In Section 3.3 we gave an account of reasoning with mental models that stressed the isomorphism between such models and the world. Crucial conclusions (e.g., about avoiding threats) can simply be read off such models without explicit inferencing. In cases of uncertainty, a single isomorphic model does not exist at the required level of specificity. Different methods exist for preserving isomorphism in this situation: for example, adopting a more generic description of the situation, or assuming that one of the possible representations is the case. Hypotheses 25 through 27 specify design concepts for dealing with ATR conclusion uncertainty through a combination of generic and assumption-based displays. Hypothesis 28 provides for the additional option of explicitly displaying alternative possibilities.

We will first address experimental tests of the basic psychological constructs, and then move to tests of the proposed ATR interface concepts.

Procedures to assess mental model reasoning: Experiments in this phase will explore the role of assumption-based reasoning when target classification is uncertain. The performance measurement paradigms described Sections 4.1.1 and 4.1.2 lend themselves naturally to this task. In Section 4.2.1 on visual processing, we discussed how these paradigms could be used to explore the probability of different kinds of confusion errors as a function of time. In that research we were most concerned with similarity parameters, $\xi(i,j)$. In this study, by contrast, our primary focus is the bias parameter, $B(i)$. The $B(i)$ determine how decision makers *make choices* about classification conclusions when the conclusions are not fully constrained by data. Consistent effects regarding these parameters indicate a systematic policy of adopting assumptions.

We will investigate (i) whether such systematic assumption-adoption policies exist, and (ii) what determines the types of assumptions that are made. In particular, we will ask whether, and under what conditions, assumptions are based on the worst-case (or best-case) classification, the most likely classification, specific categories of targets, or the classification most conducive to accomplishing a task.

To answer these questions, the experimental design will have the following features: Individual targets will be presented for classification, e.g., at the level APC, tank, truck. The data provided for some of the targets will be inconclusive regarding the required level of classification, e.g., some of the stimuli could be either APCs or tanks, either tanks or trucks, or either APCs or trucks. Several contextual variables that might affect assumptions will be varied orthogonally:

1. Mission goal in terms of specific categories of targets, e.g., kill APCs versus kill trucks versus kill tanks.
2. Mission priorities in terms of the relative costs of missing a hostile or killing a friendly, e.g., a close battle where friendlies and hostiles are mixed versus a deep battle where only hostiles are expected.
3. The relative frequency of the different possible target types (e.g., APC, truck, tank).
4. The threat to own aircraft represented by one of the target types, e.g., the range of a tank's gun.

Classification data will be analyzed (as described in Section 4.1.1) for the role of bias in confusion errors. Systematic effects on bias will then be tested for the following variables:

- Specific categories of targets: For example, do subjects tend to classify targets that might be either trucks or APCs as APCs?

- Worst-case or most dangerous: Are subjects more likely to resolve uncertainty in favor of tank in conditions where tanks are more dangerous to own aircraft than in conditions where tanks are less dangerous to own aircraft?
- Mission success: Are subjects more likely to assume that a target is a truck if the mission calls for engaging trucks, and more likely to assume the target is an APC if the mission calls for engaging APCs? Are subjects more likely to assume that a target is a friendly if the cost of killing friendlies is relatively high, and more likely to assume a target is a hostile if the cost of missing hostiles is relatively high? Are assumptions less systematic (i.e., more random) in low-stakes discriminations (e.g., between enemy APCs and trucks) than in high-stakes discriminations (e.g., between friendlies and hostiles)?
- Base rates: Are subjects more likely to classify a target as a truck in conditions where trucks are relatively more frequent than in conditions where trucks are relatively less frequent?

We will investigate the possibility of individual differences in the likelihood of each of these assumption-adoption policies. For example, more experienced operators may respond more systematically to variables such as mission success, danger, and base rate than do inexperienced subjects.

Procedures to test ATR interface hypotheses: The results of the experiments on assumption-based classification will be used to fine-tune or revise the ATR interface design hypotheses. In particular, under conditions of high stakes and high time stress, when conclusions are uncertain at the level of specificity required by the mission, Hypothesis 27 proposes that conclusions be displayed at the required level of specificity, based in part on assumptions. The experiments discussed above will determine the types of assumptions that might plausibly be utilized in these displays. The hypotheses regarding display of uncertain conclusions will be implemented in an interactive ATR mock-up and tested.

The test environment will be the engagement decision paradigm discussed in previous sections. A mission will be described, and 5 or 6 targets will be presented with the task of using the ATR displays to select the most appropriate target for engagement. Time stress will be manipulated by varying the time available for making the engagement decision (e.g., from 2 seconds to 40 seconds). The stakes of the decision will also be varied; for example:

- Low stakes: discriminating the preferred target (e.g., trucks) from among a set of acceptable hostile targets
- High stakes: discriminating the most dangerous threat in terms of its weapons range, or discriminating friends from foes.

A proportion of the ATR target classifications in each trial will be uncertain at the level required by the mission. The following ATR display conditions will be varied for these uncertain targets:

- Generic conclusions: display of a conclusion only down to the level of specificity at which the ATR is confident.
- Assumption-based conclusions: display of a conclusion at the required level of specificity based on assumptions.
- User choice: users select whether generic conclusions or assumption-based specific conclusions will be displayed.

There is as design tradeoff between the display of assumption-based versus generic conclusions. Assumption-based displays optimize the *use* of classification conclusions for engagement decisions. Generic displays optimize the opportunity to *improve* those conclusions before using them. The appropriate resolution of this tradeoff depends on the time available to make the engagement decision:

We predict that assumption-based conclusions will result in more effective engagement performance for high time-stress conditions (Hypothesis 27). Assumption-based conclusions enable operators to use the limited available time to construct a single coherent mental model of the situation. The ability to quickly visualize the battlefield may be crucial, for example, in selecting a target based on the danger posed by threat weapon range.

On the other hand, we predict that generic conclusions will lead to better performance in low time-stress conditions (Hypothesis 26). In this case, time is available for the user to make a significant contribution to genuine (rather than assumption-based) reduction of uncertainty. His independent judgment of target classification, based on his own examination of the data, will enhance overall performance, and still allow time for construction of a mental model.

Finally, in low-stakes situations, we predict that the choice of how he allocates his time can be left to the user. In these situations, the user may be in the best position to gauge the relative advantages of immediately visualizing the battlefield in order to make engagement or routing decisions, versus reducing the uncertainty of classification (Hypothesis 25).

A final orthogonal manipulation will involve the user option of viewing alternative possible conclusions at any level of specificity. Viewing alternative possibilities may help some users in low time-stress situations in their efforts to resolve uncertainty. It may also help users visualize alternative pictures of the battlefield in engagement planning (Hypothesis 28).

4.2.4 Verification Strategies

This set of experiments examines how operators sample information to accomplish difficult classifications and to verify uncertain conclusions. Experienced problem solvers in a variety of fields appear to have more effective data collection methods than less experienced problem solvers. One strategy that seasoned operators might adopt under time stress involves emphasizing breadth rather than depth, i.e., sampling a limited amount of crucial information about each target rather than full information about only some of the targets. Another strategy involves using the ATR conclusion as a starting point and selectively examining data that might pose problems for it. We will test hypotheses both about prompting users regarding ATR uncertainty (hypotheses 29 through 32) and about the way ATRs should display data (hypotheses 33 through 36).

Research will proceed in two phases: an investigation of unaided data sampling and verification strategies, and a test of the ATR design hypotheses that support those strategies. Unlike the previous topics, both phases of this research will utilize a multiple-target, engagement paradigm, rather than the single-target classification task.

Procedures for investigating verification strategies. The basic paradigm for these studies involves presentation of multiple targets, statement of a mission, and selection by the subject of the best target for engagement. Target data will not be automatically displayed. Rather, subjects will be able to request data of any sort about any target through an interactive interface. The requested data might include, for any target:

- ATR classification conclusions
- small zoomed-in video chip (for viewing overall shape and pattern of hot spots)
- large zoomed-in video chip (for viewing parts and relations among parts)
- zoomed-in video chip of objects in spatial proximity
- non-visual information about the target
- a visual template transformed to match the target (e.g., in terms of occlusion by other objects or unusual articulation or orientation of parts)

Independent variables will include the following: time available to make the engagement decision, the number of targets, the level of ATR conclusions (appropriate for the mission requirement versus more generic than the mission requirement), and features of the targets that may be known by the operator but which are not available to the ATR. Such features might be used in resolving classifications for which the ATR is uncertain. The relevant feature will be varied among the possible data sources: It might involve overall shape or hotspots, spatial relations of parts, context of other objects, non-visual information, or appearance under conditions of occlusion or unusual orientation of parts.

Two major dependent variables will be utilized: success in selecting the optimal target for engagement, and the pattern of information requests over targets. We will use the pattern of information requests to identify the data sampling strategies that subjects adopt under different conditions of workload (i.e., time and number of targets), different levels of ATR results, and different types of information available to the human. We will use engagement decision performance to evaluate the relative success or failure of the different strategies that have been identified.

Heuristic models will be developed to account for the success or failure of different sampling strategies. These models will relate performance in this paradigm to the experimental variables and to the speed-accuracy models of recognition performance developed previously. We now illustrate how such models can be used to generate predictions and explain findings.

Imagine, for example, a scenario with four targets. The mission is to engage the greatest threat. Each target is a hostile tank, but one is the most dangerous because it has a long-range gun while the others have short-range guns. Thus, the relevant level of classification is the specific type of tank, since this is correlated with degree of threat.

The probability, *E(T)*, that the observer will select the best target for engagement after *T* seconds in the engagement task, depends on the probability, *C(t)*, that the observer will classify a given target correctly at the relevant level of specificity, after t seconds of focusing on that individual target. The exact relationship between *E* and *C* depends on the mapping from *T* to *t*: How much time, *t*, will the subject spend on an individual target given that he has a total of T seconds to perform the engagement task? This mapping, in turn, will depend on the strategy observers adopt for sampling information about targets within the allotted time, *T*. For illustrative purposes, we will discuss two kinds of strategy: information-centered and target-centered.

Information-centered strategy: The observer first examines a particular type of information about each target, then examines another type of information about each target, and so on. This strategy results in an approximately equal amount of information being examined about each target. But under time constraints, it can result in the total omission of some types of information. An example of this strategy would be consulting the full screen view of each target, then (if there is time) going back to look at a zoomed-in video chip of each target, then (if there is more time) going back to look at non-visual information about each target, etc. In an ATR context, a variant of this would involve first scanning the ATR classification of each target, then (if there is time) going back to look at the video chip, then looking at non-visual information, and so on.

With the information-centered strategy, the time spent per target, *t,* would be expected (to a first approximation) to be proportional to *T/N*, i.e., the total time for the engagement task divided by the number of targets:

Eq. 1 $$t = q(T-a)/N.$$

The proportionality allows for allocation of some of *T* (represented by *a*) to other tasks: e.g., constructing a mental model that compares the different targets in terms of threat. It also allows for an inefficiency or overhead in processing targets (represented by the slope $q <= 1.0$). Such inefficiency might be caused by the more stressful engagement-decision context as compared with the simpler classification task. A more interesting potential source of inefficiency involves display design. Presentation of large quantities of less useful information may impede the operator's ability to extract the information he needs.

Suppose that at a given time *T* into the engagement task, the observer has spent the equivalent of *t* seconds observing each target. Speed-accuracy operating characteristics for classification indicate that the expected classification accuracy for an individual target, based on *t* seconds of processing, is C_t. (I.e., there is chance C_t that the operator will classify a long-range weapon platform as a long-range platform, and a chance C_t that he will classify a short-range weapon platform as a short-range platform). What is the probability that this operator will correctly chose the best target for engagement?

The probability, *E(T)*, of selecting the best target for engagement is a function of (i) accurate classification of the best target and (ii) possible inaccurate classification of one or more of the inappropriate targets, plus guessing:

Eq. 2

$$E(T) = \sum_{k=0}^{N-1} \binom{N-1}{k} C_t^{N-k} (1-C_t)^k \left(\frac{1}{k+1}\right)$$

In equation 2, C_t represents the probability of accurately classifying any *individual* target in the respect that is relevant for engagement, based on the information that is sampled about that target in time *t* (*t* is related to *T* by equation 1 above). The equation makes the simplifying assumption that if more than one target is classified as appropriate for engagement, the observer selects among them randomly.[2]

As the number of targets increases, engagement performance declines for two separate reasons: First, classification accuracy C_t decreases as the time allocated to each target declines. In addition, however, the chance of at least one classification error rises simply because there are a larger number of targets (even if C_t remained constant). For example, if adequate time is available for fully processing a single target, suppose individual-target classification accuracy is 90%. With four targets, even if C_t stays at 90%, the probability of making the right engagement choice *E(T)* falls to 77%. Suppose (more plausibly) that when four targets are presented, the probability of correctly classifying an individual target falls to 80%. Then the probability of making the right engagement choice is only .49.

Target-centered strategy. Another strategy that observers might adopt is target-centered. In this strategy, the observer examines all the information about a target before going on to examine all the information about the next target, and so on. Under time constraints, this strategy can result in the complete omission of some targets.

[2] $C^{N-1-k}(1-C)^k$ represents the probability of one way that *k* of the *N-1* targets that are not in fact appropriate could be misclassified as appropriate for engagement. The best target can be selected only if it is classified as appropriate for engagement, so Equation 2 assumes that the best target is correctly classified. Thus, $C^{N-1-k}(1-C)^k$ is multiplied by C, resulting in the expression $C^{N-k}(1-C)^k$ in each term. The observer guesses from among k incorrectly classified targets and 1 correctly classified target. 1/(k+1) thus represents the chance that he will pick the actual best target from this set.

The equation assumes that the operator will not engage if no target is classified as meeting the highest standard for engagement. This would happen if the best target is misclassified (as inappropriate for engagement), and all the other targets are correctly classified (as inappropriate for engagement). This assumption is probably appropriate for the hostile/friendly engagement criterion: If every target is classified friendly, there is no engagement. But in other situations (e.g., killing the greatest threat among a set of hostiles), the operator might randomly select a target for engagement from among all the targets when one appears to be no greater threat than the others. In that case, a term must be added to the right side of equation 2 to represent the chance of accidentally selecting the best target from the whole set:

$C^{N-1}(1-C)(1/N)$

With the target-centered strategy, the number of targets, *n*, that an observer is able to examine depends on *T*, the amount of time available for the engagement decision, and on *t**, the amount of time it takes to fully process the information provided for an individual target (i.e., to achieve asymptotic classification accuracy):

Eq. 3 $$n = \min(N, q(T-a)/t^*).$$

Here again, *a* represents time that might be allocated to other activities than classifying individual targets, and *q<=1.0* represents a potential inefficiency in using the time that is left.

With this strategy, there is a major difference between performance with targets that have been examined and targets that have not been examined. If a target has been examined, the expected classification accuracy equals the maximum (or asymptotic) classification accuracy, C_{t*}, achievable with the given information. If a target has not been examined, we assume that it will not be engaged.

Suppose the best target for engagement happens to be in the examined set. Then the probability of correctly choosing it after time *T*, *E*(T)*, is given by equation 2 for the information-centered strategy, but substituting C_{t*} for C_t, and substituting the actual number of targets examined in the available time (*n* as a function of *T,* according to equation 3) in place of the total number of targets, *N*:

Eq. 4 $$E^*(T) = \sum_{k=0}^{n-1} \binom{n-1}{k} C_{t^*}^{\,n-k} (1 - C_{t^*})^k \left[\frac{1}{k+1}\right]$$

But the best target may not be in the examined set. This strategy assumes that there is no dependable prior information by means of which the operator can select the most promising targets for examination. Thus, the best target has approximately the same chance of being examined as any other target, and the chance that the examined set of targets includes the best target is *n/N*. The final probability, *E'(T)*, of selecting the best target for engagement is just the chance of both examining the best target and selecting it from the examined set:

Eq. 5 $$E'(T) = E^*(T)\, n/N.$$

The target-centered strategy degrades rapidly as the number of targets increases. *E*(T)* is not affected (since it assumes that the best target is examined), but the chance of examining the best target, *n/N,* will decline rapidly. Suppose, as in our previous example, that only a single target is presented, and there is adequate time for processing it; individual-target classification accuracy *E*(T)* is 90%. If four targets are presented, *E'(T)* falls to 24%.

The relative advantages of the information-centered and target-centered strategies may depend on the circumstances. The information-centered approach will be better whenever *E(T)* is a bigger proportion of *E*(T)* than *n* is of *N*: i.e.*,*

$$E/E^* > n/N$$

from equations 2 and 5 (omitting the argument *T* for convenience). Under time stress, e.g., for small values of *T* or large values of N, the information-centered strategy has an advantage: The initially sampled data for each target will generally be the most valuable, and thus *E* will be relatively large; on the other hand, because of the limited time, or the large number of potential targets, *n/N* will be small. On the other hand, in non-stressed situations, i.e., for large values of *T* or small values of *N*, *E* converges on *E** and *n/N* converges on 1; thus, the two approaches might tend to become equivalent.

Why would the target-centered strategy ever be used? One reason could be processing efficiency, reflected in the parameters *a* and *q* in equations 1 and 3. Unlike the information-centered strategy, the target-centered strategy permits a complete evaluation of each target before going on to the next one. All that need be retained in memory is the result of classifying each examined target (appropriate or inappropriate for engagement). The information-centered strategy, by contrast, requires retention of a running summary of the evidence for each target through repeated iterations. We thus predict greater use of information-centered strategies for time-stressed conditions, and greater use of target-centered strategies for non-time-stressed conditions.

In fact, operators might not use a a pure target-centered strategy very often. They will not ususally have to select targets randomly for examination without any prior indication of their relative value. Graphical displays, for example, place all targets in a common spatially represented situation. At a minimum, then, common information is generally available across all targets about their spatial locations relative to one's own aircraft, and their spatial relations to one another. Such information enables the operator to crudely prioritize the targets for examination. A more effective variant of this strategy, then, might be a hybrid, in which the same basic information is first extracted about all targets, then (if time permits), the most significant targets are subjected to a more exhaustive examination.

A particularly interesting hybrid strategy is possible when the user is interacting with an ATR. In the selective critiquing strategy, the observer first examines the ATR conclusion regarding each target (an information-centered approach), then proceeds to examine additional information about targets whose ATR classification seems questionable in respects relevant for engagement (a target-centered approach). This strategy usually results in sampling at least one piece of information on each target (the ATR classification), but can result in no further information about some targets.

We predict that more experienced operators will adopt the selective critiquing strategy. Its advantages are: (a) the ATR classification is a relatively valuable piece of information, and all targets are initially sampled with respect to that information; and (b) complete information is efficiently sampled about targets where the user's attention can do the most good. This strategy involves some risk-taking: i.e., that the operator can intelligently select targets for detailed examination. A bad choice could result in no additional attention paid to important targets, e.g., friends about to be engaged or threats about to be missed. Some predictions in this respect are possible: Experienced operators will be highly adaptive to the demands of a particular situation in their use of this strategy. They will be influenced by mission priorities in their selection of targets for detailed attention, choosing targets that involve potential false alarms when the cost of killing a non-target exceeds the cost of missing a foe and choosing targets that involve potential misses when the opposite is the case. Moreover, in sampling additional information, experienced operators will give priority to types of information that are relevant to classification but not processed by the ATR.

We predict that the skills of relatively inexperienced operators will be less developed in both these respects. Inexperienced decision makers tend to adopt fixed procedures rather than adapt their strategies to the particular demands of a situation. We thus predict that inexperienced operators will be less likely to use a selective critiquing strategy, and more likely to adopt the less risky information-centered strategy, in which some information is collected about all targets.

Procedures for testing ATR interface design concepts: In the next phase of experimentation, we will test the ability of the ATR interface to support efficient user verification strategies. The same engagement decision task will be utilized, but additional features of the ATR interface will be introduced as independent variables. The effects of these features will be examined under different conditions of stakes and time stress.

The following variables will be investigated for conditions where the ATR is significantly uncertain:

- Prompting versus no prompting. Prompts will be developed based on the likelihood and cost of recognition errors. Thus, in the prompting condition, prompts may alert users to potential misses, and in other cases alert them to potential false alarms.
- Display of all signals versus full information versus supporting data. In the all-signals condition, the ATR determines if there is any information content in a given category of information before displaying it. For example, if the parts of the object are discernible, a large-scale video chip will be displayed; if any objects are in spatial proximity to the target, a view of the surrounding objects will be displayed. The ATR does not evaluate the importance of the information, leaving that to the operator. In the full-information condition, all categories of information are displayed regardless of information content. In the supporting-data condition, only information used by the ATR to arrive at its conclusion are displayed.
- Mission priorities, emphasizing cost of false alarms (e.g., friendlies and hostiles mixed in a close battle) versus emphasizing cost of misses (e.g., deep battle with only hostiles).
- Stakes, or costs of errors (e.g., the potential danger of a missed threat, the cost of a killed friend)
- Time to accomplish engagement decision (e.g., 2 to 40 seconds).

Hypotheses 31 and 33 predict that ATR prompts and the all-signals display condition will combine to facilitate use of an efficient and adaptive selective-critiquing strategy under conditions where there is adequate time to carry it out (low time stress) and where the costs of errors make examination of additional data worthwhile (moderate to high stakes). Hypothesis 32 and 33 predict that the selective-critiquing strategy may also be valuable under high time stress, in conditions where the stakes of an error are high enough that it is worth the chance of missing a potential target. Finally, Hypothesis 34 suggests that supporting data only should be displayed with low stakes or high time stress and moderate stakes.

Subsidiary manipulations might include comparison of numerical uncertainty indicators with binary indicators (uncertain versus certain). Hypothesis 29 predicts that binary prompts will lead to more efficient engagement decision making. We might also compare various ways of symbolizing uncertainty in the same context (Hypothesis 30).

4.3 Tasks and Schedule

The proposed research will be carried out in four sets of studies. Experimental sets 1 and 2 will be performed in the first 15 months of the project, and sets 3 and 4 will be completed in the second year:

Set 1: Includes the basic psychological research on stages of visual processing (Section 4.2.1) and on levels of categorization (Section 4.2.2).

Set 2: Includes tests of the ATR interface design hypotheses regarding display of data (Section 4.2.1) and ATR conclusions (Section 4.2.2).

Set 3: Includes the basic psychological research on mental models in reasoning about uncertainty (Section 4.2.3) and on verification strategies (Section 4.2.4).

Set 4: Includes tests of ATR design hypotheses regarding display of uncertain ATR conclusions Section 4.2.3) and display of prompts and data under uncertainty (Section 4.2.4).

Note that each year involves a cycle of research beginning with basic psychological processes that have implications for ATR design (sets 1 and 3), and proceeding to the ATR design implications of those processes (sets 2 and 4). The first year (sets 1 and 2) focuses on more elementary processes (visual recognition and verbal categorization) and their implications, while the second year (sets 3 and 4) examines the consequences of introducing uncertainty, i.e., the more advanced processes of reasoning with mental models and verifying ATR conclusions. Figure 8 diagrams some of the relationships among the sets of studies.

Each set of studies will embed four subtasks:

1. Detailed experimental design
2. Implementation of experimental environment (preparation of stimuli and programming of software)
3. Data collection
4. Analysis and Conclusions

Figure 9 provides a complete schedule of tasks and deliverables.

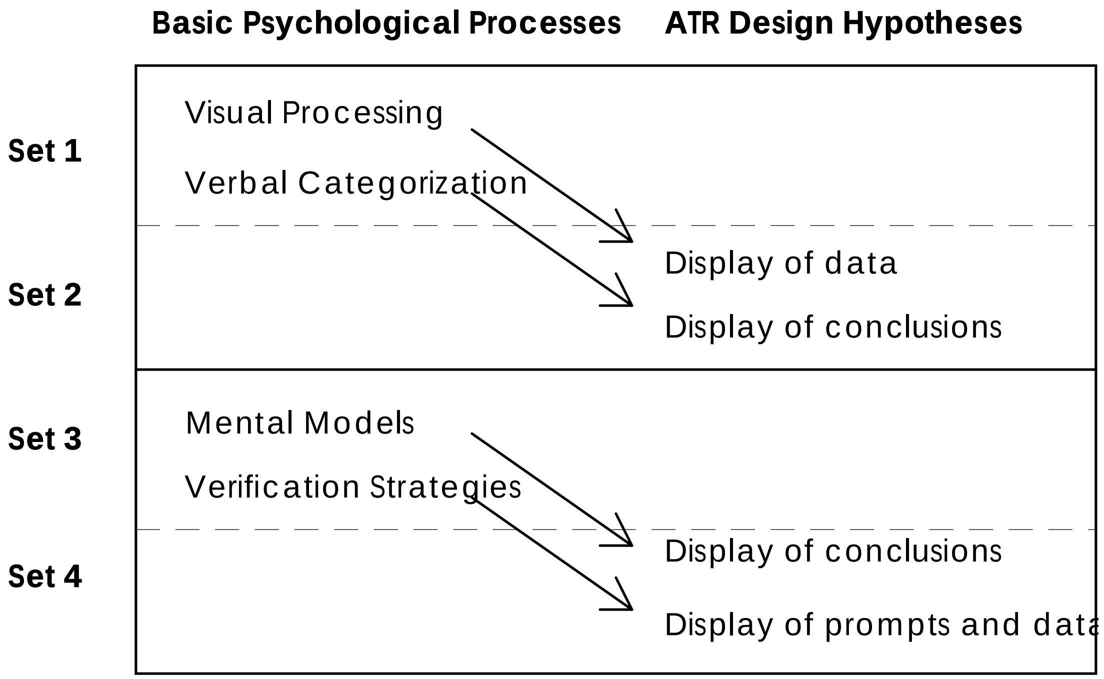

Figure 9. Relationships among experimental sets.

4.4 Subjects and Equipment

Experiments will be run, as much as possible, with active-duty Army helicopter pilots. All subjects will all have experience in tasks involving target recognition, although the degree of experience may vary from subject to subject.

Subjects will be obtained at Fort Bragg, NC, and at Fort Hood, TX, through cooperative relationships we have established with researchers at the U.S. Army Night Vision and Electro-Optics Directorate, Fort Belvoir, VA. Brian Gillespie has been an extremely helpful point of contact at NVEOD for arranging subjects. LTC Gary Botts was our point of contact at Fort Bragg for the Phase I interviews.

Experiments will be implemented on an IBM desktop computer and high-resolution color monitor. A likely software environment is Windows and C.

We will attempt to use realistic image stimuli to the greatest extent possible, within the constraints of the experimental designs. Realistic stimuli with a great variety of salient features and degrees of uncertainty are available from NVEOD. We have ascertained that these stimuli can be displayed on CTI computer equipment. Clare Walters is our point of contact.

Figure 9 here

REFERENCES

Adelson, B. When novices surpass experts: The difficulty of a task may increase with expertise. Journal of Experimental Psychology: Learning, Memory, and Cognition, 1984, 10(3), 484-495.

Anderson, J.R. The adaptive character of thought. Hillsdale, NJ: Lawrence Erlbaum Associates, 1990.

Anglin, J.M. Extensional aspects of the preschool child's word concepts. In T.B. Seiler and W. Wannenmacher (Eds.), Concept development and the development of word meaning. New York: Springer-Verlag, 1983, 247-266.

Barsalou, L.W. Deriving categories to achieve goals. In G.H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Volume 27). New York: Academic Press, 1991, 1-64.

Barsalou, L.W. Cognitive psychology: An overview for cognitive scientists. Hillsdale, NJ: Lawrence Erlbaum Associates, 1992.

Becker, C.A., Hayes, B.C., and Gorman, P.C. User acceptance of intelligent avionics: A study of automatic-aided target recognition. Eugene, OR: Bio-Dynamics Research and Development Corp., October 1990.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P. Discrete multivariate analysis: Theory and practice. Cambridge, MA: MIT Press, 1975.

Bliss, W. A review of the literature on briefing and target acquisition performance (NWC TP 5650). China Lake, CA: Naval Weapons Center, 1974.

Broadbent, D.E. Decision and stress. New York: Academic Press, 1971.

Brown, A.L., and DeLoache, J.S. Skills, plans, and self-regulation. In R.S. Siegler (Ed.), Children's thinking: What develops? Hillsdale, NJ: Lawrence Erlbaum Associates, 1978.

Byrne, R.M.J. The model theory of deduction (Chapter 2). In Y. Rogers, A. Rutherford, and P.A. Bibby (Eds.), Models in the mind: Theory, perspective and application. New York: Academic Press, 1992.

Chase, W.G., and Simon, H.A. Perception in chess. Cognitive Psychology, 1973, 4.

Chi, M., Feltovich, P., and Glaser, R. Categorization and representation of physics problems by experts and novices. Cognitive Science, 1981, 5, 121-152.

Chi, M., Glaser, R., and Rees, E. Expertise in problem solving. In R.S. Steinberg (Ed.), Advances in the psychology of human intelligence (Vol. 1). Hillsdale, NJ: Lawrence Erlbaum Associates, 1982, 7-75.

Chinnis, J.O., Jr., Cohen, M.S., and Bresnick, T.A. Human and computer task allocation in air-defense systems (Technical Report 84-2). Falls Church, VA: Decision Science Consortium, Inc., August 1984.

Cohen, M.S. Mental models, uncertainty, and in-flight threat responses by air force pilots. Proceedings of the Fourth Biennial Symposium on Aviation Psychology. Columbus, Ohio, April 1987.

Cohen, M.S. Decision making "biases" and support for assumption-based higher-order reasoning. To appear in the Proceedings of the Fifth Workshop on Uncertainty and AI, Windsor, Ontario, August 1989.

Cohen, M.S. Taking risks and taking advice: The role of experience in airline pilot diversion decisions (Draft Final Technical Report). Fairfax, VA: Decision Science Consortium, Inc., June 1992.

Cohen, M.S. Three paradigms for viewing decision biases. The naturalistic basis of decision biases. The bottom line: Naturalistic decision aiding. In Klein, Orisanu, and Calderwood (Eds.), Decision making in complex worlds. In press.

Cohen, M.S., Brown, R.V., Seaver, D.A., and Ulvila, J.W. Operability in attack submarine combat systems: An exploratory overview (Technical Report 82-2). Falls Church, VA: Decision Science Consortium, Inc., April 1982.

Cohen, M.S., Leddo, J.M., and Tolcott, M.A. Cognitive strategies and adaptive aiding principles in submarine command decision making. Submitted to Management Science, October 1988.

Cohen M.S., Leddo, J.M., and Tolcott, M.A. Personalized and prescriptive aids for commercial air flight replanning (Technical Report 89-2). Reston, VA: Decision Science Consortium, Inc., April 1989.

Cohen, M.S., Tolcott, M.A., and McIntyre, J.R. Display techniques for pilot interactions with intelligent avionics: A cognitive approach (Technical Report 87-6). Falls Church, VA: Decision Science Consortium, Inc., April 1987.

Cohen, M.S., Adelman, L., Leddo, J.M., Tolcott, M.A., and Chinnis, J.O., Jr. Human and computer task allocation in air defense systems: Phase II annual report (Technical Report 87-2). Falls Church, VA: Decision Science Consortium, Inc., April 1987.

Connolly, T., and Wagner, W.G. Decision cycles. Advances in information processing in organizations, 1988, 3, 183-205.

Corter, J.E., and Gluck, M.A. Explaining basic categories: Feature predictability and information. Psychological Bulletin, 1992, 111(2), 291-303.

Cruse, D.A. The pragmatics of lexical specificity. Journal of Linguistics, 1977, 13, 153-164.

Curry, R.E. The introduction of new cockpit technology: A human factors study (NASA Technical Memorandum 86659). Moffett Field, CA: NASA-Ames Research Center, 1985.

Defense News. Pentagon seeks second look at combat identification needs, March 25, 1991.

Doyle, J. A truth maintenance system. Artificial Intelligence, 1979, 12, 231-272.

Einhorn, H.J. Learning from experience and suboptimal rules in decision making. In T.S. Wallsten (Ed.), Cognitive processes in choice and decision behavior. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

Fitts, P.M. (Ed). Human engineering for an effective air-navigation and traffic-control system. Columbus, OH: Ohio State University Research Foundation, 1951.

Glaser, R. Expertise and learning: How do we think about instructional processes now that we have discovered knowledge structures? (Chapter 10). In D. Klahr and K. Kotovsky (Eds.), Complex information processing: The impact of Herbert A. Simon. Hillsdale, NJ: Lawrence Erlbaum Associates, 1989, 269-282.

Green, D.M., and Swets, J.A. Signal detection theory and psychophysics. New York: John Wiley and Sons, Inc., 1966.

Hammond, K.R., Hamm, R.M., Grassia, J., and Pearson, T. The relative efficacy of intuitive and analytical cognition: A second direct comparison (Report No. 252). Boulder, CO: University of Colorado, Center for Research on Judgment and Policy, June 1984.

Huber, G.P. Cognitive style as a basis for MIS and DSS designs: Much ado about nothing?. Madison, WI: University of Wisconsin, 1982.

Johnson-Laird, P.N. Mental models: Towards a cognitive science of language, inference, and consciousness. Cambridge, MA: Harvard University Press, 1983.

Johnson-Laird, P.N., and Byrne, R.M.J. Deduction. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.

Joliceur, P., Gluck, M., and Kosslyn, S.M. Picture and names: Making the connection. Cognitive Psychology, 1984, 16, 243-275.

Kadane, J.B., and Lichtenstein, S. A subjectivist view of calibration (Report 82-6). Eugene, OR: Decision Research, 1982.

Kosslyn, S.M. Image and mind. Cambridge, MA: Harvard University Press, 1980.

Kosslyn, S.M., and Koenig, O. Wet mind: The new cognitive neuroscience. New York: The Free Press, 1992.

Larkin, J.H. Problem solving in physics (Technical Report). Berkeley: University of California, Group in Science and Mathematics Education, 1977.

Larkin, J.L. Enriching formal knowledge: A model for learning to solve textbook physics problems. In J.R. Anderson (Ed.), Cognitive skills and their acquisition. Hillsdale, NJ: Lawrence Erlbaum Associates, 1981.

Larkin, J., McDermott, J., Simon, D.P., and Simon, H.A. Expert and novice performance in solving physics problems. Science, June 1980, 208, 1335-1342.

Lehner, P.E., Cohen, M.S., Mullin, T.M., Thompson, B.B., and Laskey, K.B. Adaptive decision aiding (Technical Report 87-3). Falls Church, VA: Decision Science Consortium, Inc., February 1987.

Lehner, P.E., Mullin, T.M., and Cohen, M.S. Adaptive decision aids: Using fallible algorithms to support decision making. Submitted to Management Science, 1988.

Libby, R., and Lewis, B.L. Human information processing research in accounting: The state of the art. Accounting, Organizations and Society, 1977, 2(3), 245-268.

Lowe, D.G. Three-dimensional object recognition from single two-dimensional images. Artificial Intelligence, 1987, 31, 355-395.

Luce, R.D. Individual choice behavior. New York: Wiley, 1959.

Luce, R.D. The choice axiom after twenty years. Journal of Mathematical Psychology, June 1977, 15(3).

MacMillan, J. Plan for experiments on alternative human-computer interfaces for machine-aided target acquisition (TR-537). Burlington, MA: Alphatech, Inc., 5 March 1992.

Navon, D. Forest before trees: The precedence of global features in visual perception. Cognitive Psychology, 1977, 9, 353-383.

Neisser, U. Cognitive psychology. New York: Appleton-Century-Crofts, Meredith Corporation, 1967.

Neisser, U. Cognition and reality. Principles and implications of cognitive psychology. San Fransciso: W.H. Freeman and Company, 1976.

Newell, A., and Simon, H.A. Human problem solving. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Pachella, R.G., and Pew, R.W. Speed-accuracy tradeoff in reaction time: The effect of discrete criterion times. Journal of Experimental Psychology, 1968, 76, 19-24.

Pachella, R.G., Smith, J.E.K., and Stanovich, K.E. Qualitative error analysis and speeded classification. In J. Castellan (Ed.), Cognitive Science III. Hillsdale, NJ: Lawrence Erlbaum Associates, 1978, 169-198.

Palmer, S.E. The effects of contextual scenes on the identification of objects. Memory and Cognition, 1975, 3, 519-526.

Patel, V.L., and Groen, G.J. The general and specific nature of medical expertise: A critical look. In K.A. Ericsson and J. Smith (Eds.), Toward a general theory of expertise: Prospects and limits. Cambridge: Cambridge University Press, 1991.

Payne, J.W. Task complexity and contingent processing in decision making: An information search and protocol analysis. Organizational Behavior and Human Performance, 1976, 16, 366-387.

Reed, A.V. List length and the time course of recognition in immediate memory. Memory and Cognition, 1976, 4, 16-30.

Reiter, R. A logic for default reasoning. Artificial Intelligence, 1980, 13, 81-132.

Riley, M.S., Greeno, J.G., and Heller, J.I. Development of children's problem-solving ability in arithmetic. In H.P. Ginsburg (Ed.), The development of mathematical thinking. New York: Academic Press, 1983, 153-196.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., and Boyes-Braem, P. Basic objects in natural categories. Cognitive Psychology, 1976, 8, 382-439.

Russo, J.E., and Dosher, B.A. Cognitive effort and strategy selection in binary choice. University of Chicago, Graduate School of Business and Columbia University, May 1981.

Sage, A.P. Behavioral and organizational consideration in the design of information systems and processes for planning and decision support. IEEE Transactions on Systems, Man, and Cybernetics, 1981, 11(9), 640-678.

Schoenfeld, A.H., and Herrman, D.J. Problem perception and knowledge structure in expert and novice mathematical problem solvers. Journal of Experimental Psychology: Learning, Memory, and Cognition, 1982, 8, 484-494.

Sheridan, T.B. Supervisory control. In G. Salvendy (Ed.), Handbook of human factors/ergonomics. New York: Wiley, 1987.

Svenson, O. Process descriptions of decision making. Organizational Behavior and Human Performance, 1979, 23, 86-112.

Toms, M.L., and Kuperman, G.G. Sensor fusion: A human factors perspective. Dayton, OH: Logicon Technical Services Incorporated, September 1991.

Townsend, J.T. Theoretical analysis of an alphabetic confusion matrix. Perception and Psychophysics, 1971, 9, 40-50.

Weiser, M., and Shertz, J. Programming problem representation in novice and expert programmers. International Journal of Man-Machine Studies, 1983, 19.

Wiener, E.L. Human factors of cockpit automation: A field study of flight crew transition (NASA Contractor Report 177333). Coral Gables, FL: University of Miami, 1985.

Wiener, E.L. Cockpit automation. In E.L. Wiener and D.C. Nagel (Eds.), Human factors in aviation. San Diego: Academic Press, 1988.

Wright, P., and Barbour, F. Phased decision strategies: Sequels to an initial screening. In M.K. Starr and M. Zeleny (Eds.), Multiple criteria decision making. Amsterdam: North-Holland, 1977.