1

ADVANCED DECISION SYSTEMS
1500 Plymouth Street, Mountain View, California 94043-1230 • (415) 960-7300 • FAX (415) 960-7500

Counter Deception

Gregory J. Courand

December 1, 1989

Progress Report
Contract No: N00014-89-C-0299
Period Covered: September 1 1989 — November 30, 1989
ADS Project: 3263

Prepared for: Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217

The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Advanced Decision Systems position, policy or decision unless so designated by official documen ion.

CONTENTS

LIST OF FIGURES

1. Introduction

In this report we describe our research on the problem of providing counter-deception support for expert systems. This report represents our views at the midpoint of the contract, organized as a complete but not-fully-worked out final report. We define the problem, present an overview of the technical challenges that we believe must be addressed, and conclude with an introduction to three sample domains that could profit from the support we envision. Since this first-phase SBIR project is so brief we plan to spend the remainder of the project elaborating the technical agenda and will seek to apply it to the sample domains.

An expert system is deceived only if its product (e.g., a situation assessment) contains errors that derive from intentional actions of the enemy to mislead the system and which give the enemy an advantage. The possibility of deceiving these systems constitutes a serious threat to national security, particularly as more and more expert systems are being fielded to support military planning, intelligence analysis, and decision-making, The government speculates that it may be relatively easy to deceive some expert systems, Hence there is a need for basic research on the possibility of strengthening expert systems so that they are less sensitive to deception.

We concur with the government: deception of expert systems is a serious and growing threat. On military grounds, it is not hard to see the value of speculating about the information-processing capabilities of one's enemy, looking for weaknesses and omissions that can be exploited. Classically, this value rises as the stakes of the conflict rise.[1] If the enemy can acquire a copy of the expert system then his problem is that much simpler; however, he does not need a copy to be an effective deceiver. On technical grounds, it is clear that the overwhelming thrust of the research in expert systems has been to faithfully and effectively compute symbolic expressions that plausibly or logically follow from the information and goals provided to the system. If the information has been suitably manipulated then the system will faithfully and effectively lead itself down the garden path. Ultimately, the technical problem is that the expert system lacks an explicit representation for deception and thus has no (computational) basis to examine its conclusions in this light.

[1] It is interesting and important to note in this regard that the Soviets treat "reverse engineering" as a prominent engineering discipline, almost as a science. They are the world leaders in this practice. It would be foolish to believe that they practice reverse engineering on electro-mechanical but not knowledge-based systems.

We believe that deceptive efforts actually target not only the expert system but also the sensor suite that provides data for the expert system and the human analyst, planner, or decision maker that uses the tool. It is this larger system, comprised of the sensors, expert system, and human users that must be supported. Our charter is to augment expert systems, not sensor systems or human beings per se, but this does not rule out supporting the larger system. Therefore we define our primary goal as: augment the expert system with a representation of deception. The representation may be internal to the expert system (e.g., some behavior models can be marked as 'deceptive' and then used accordingly). Alternately, the representation can be external to the system (e.g., tools to modify modules of the expert system and/or partition the data of the system to account for deception). Thus we provide direct support for counter-deception. Our secondary goal is to endow the expert system with the ability to reason about the strengths and weaknesses of the sensor systems that 'feed' it. Specifically, we plan to augment the expert system with mechanisms so that it can reason about whether certain mechanics of the deception could be carried out without being observed. Our final goal, resources permitting, is to find a way to account for the cognitive factors that underly the way a human analyst uses the tool. Our view is that cognitive factors manifest themselves as a commitment to one set of beliefs over another that extends beyond the (rational) evidential support for those beliefs. We suspect that maintenance and review of the reasons for system conclusions may augment plausibility measures in a way that helps to expose and limit cognitive biases — among other things, the presence of reasons delimits what has and has not been considered.

We propose three types of support for counter-deception. First, the analyst must be able to explore the possibility that some of the behaviors that have been observed are meant to deceive. Tools are built so that the analyst can test whether observations may be the result of deceptive behaviors and if so derive the implications of the underlying/true state. Specifically, the analyst constructs new behavior models that yield the actual observations but contain deceptive behaviors as the basis for some of the observations. Once a feasible new behavior model is created it is analyzed to determine its requirements, the possible goals it serves, and the implications the underlying/true state has for us. For example, it may be possible, using decoys, to maintain two Typhoon-class SSBNs in the time we currently think it takes to maintain one. This has implications for their maintenance capability and for the required fidelity of the decoy, may illuminate their goals, and (once their readiness given double-maintenance is computed) may undermine our view of readiness parity. The other types of support augment this model-building and exploring facility. Design principles must be enunciated so that expert systems will be inherently more resistant to deception. Also, tools that reside in the (embedding) environment of the expert system are required: e.g., to execute the expert system under different conditions and compare the results.

The type of support we are proposing bears emphasis. We are not proposing automated detection of deception; this is not feasible. We do not believe that the answer lies solely with proper use of an evidential calculus: after all, a good deception is very plausibly received. And we have decided that reasoning at length about the beliefs agents have about the beliefs of other agents is not powerful enough to serve as the basis for counter-deception support.[2] Rather, we are proposing that the armamentarium of the counter-deception analyst primarily contains deceptive concepts and techniques and the ability to integrate these into honest (i.e., non-deceptive) behaviors. This calls for model-based and constraint-based tools that can elaborate the structures employed by the expert system. At a deeper level, it calls for a notion of plausibility that applies to models, not just data.

The remainder of the report is organized as follows. Chapter 2 describes the deception problem, includes an example of deception for SSBN maintenance, and presents our technical proposals to support counter-deception. Chapter 3 contains a collection of issues that we plan to explore in the upcoming weeks, along with a few sample domains we could use to test our concepts and tools. Appendix A describes the representation of belief and shows how deception leads to certain systematically-maintained discrepancies between the beliefs of deceiver and deceived. Finally, Appendix B contains a collection of references that we have found to be pertinent to our investigations.

[2] However, evidential calculi and the language of belief relations will have a role. Also, we have included an appendix that describes the role of belief relations.

2. Technical Investigations

Deception is a behavior that strives to create and maintain a sequence of false states of belief in the enemy. It is rarely just a false object (e.g., a decoy) or configuration. In dramatic terms, it is a play enacted with leading and supporting roles; it is a process rather than a set-piece — and the final act violates expectations. For example, operation "Fortitude" was responsible for deceiving Germany into believing that the Allied invasion would take place at the Pas de Calais, 200 miles east of Normandy.[1] It was very clear to the designers of Fortitude that it would not be enough to build intelligence and support facilities (i.e., buildings and other physical structures) to suggest the attack at the Pas de Calais. Instead, radio traffic was routed through these facilities and reconnaissance missions were flown from this site. Most important, an entire army group formation (FUSAG, for First United States Army Group) was invented. Increasingly detailed order of battle documents were created for FUSAG and leaked through double agents, so that German intelligence would be lead to believe, in an evolutionary and event-sensitive manner, in the existence and apparent intent of this notional force.

This suggests very directly that our technical work must address behavioral processes and, for the case of counter-deception, alternative explanations of observed processes. We begin by defining what we mean by "expert system" and then by "deception" in this report. Then we present an example of the sort of behavioral analysis we are interested in supporting. Next we identify the primary technical requirements for creating and analyzing behavioral processes that contain deceptive elements. This will lead to specific proposals for counter-deception.

2.1 Deception of Expert Systems

For our purposes, an ***expert system*** is any computational tool that constructs symbolic conclusions based on symbolic and/or numerical input and that can present a trace of the processing that lead to the conclusion. This definition is purposely very broad. We do not distinguish among architectures (e.g., centralized versus distributed), knowledge representations (e.g., frames versus quantified sentences), processing styles (e.g., logical deduction versus blackboards), or evidential calculi (e.g., ad hoc schemes versus Bayesian probability). We also do not distinguish

[1] Fortitude was a part of the Europe-wide deception operation known as "Bodyguard." Numerous references, including original source material, exist on this operation. For example, see the Cruickshank and Brown references. The original material is contained in Hesketh, "*Fortitude*: A History of Strategic Deception in North Western Europe, April 1943 to May 1945." An excerpt can be found in Chapter 10 of the Daniel and Herbig reference.

expert systems according to their 'purpose' (e.g., assessment versus planning). We do rule out purely neural or non-representational systems, since these do not (at this time) produce a trace of their reasoning. However, we permit hybrid expert systems that contain neural components.

We say that *an expert system is deceived* only if the following conditions are met:

1. the system creates its product,

2. this product contains errors which:

   a. derive from intentional actions of the enemy, and

   b. give the enemy an advantage in terms of our responses.

The main implications of this definition are as follows. First, the system creates its product, be it an assessment of a situation, a proposed plan, or a recommended decision. This is an essential part of the definition since it distinguishes deceiving the expert system from "defeating" it. To deceive the system is to behave such that our observations will be computationally tractable (i.e., likely to be 'figured out' by the expert system) and the resultant beliefs will be wrong in a way that serves the deceiver. In contrast, to defeat an expert system is to behave in a way that yields a computationally intractable problem for the system.[2] Second, the product contains errors that are the results of intentional enemy actions to mislead us. Thus mistakes made by the system and design biases in the system (that favor the enemy), while serious, are distinguished from deception. Also, we note that deception is a process and hence we refer to enemy actions rather than 'enemy states.' Third, advantage is measured in terms of our responses. This follows from the obvious fact that deception is worthless if its presence or absence elicits the same response from us (or if it elicits the 'wrong' response from us).

In the remainder of this document we do not concern ourselves with the specific product of the expert system. However, this needs to be considered in the future, since the way that deception affects an expert system depends in essential ways on whether the purpose of the system is assessment, planning, decision-making, etc.

[2] This distinction did not occur to us until we began to consider how we might deceive specific systems. Then we realized an enemy could attack an expert system by neutralizing/defeating it rather than fooling/deceiving it. Systems that face large combinatoric spaces may be fairly easy defeated. We have not seen this distinction mentioned in the deception literature and we suspect it has not been, since the problem is more likely to arise (and can be stated with greater precision) for an expert system than for a human analyst. Working with a colleague, Tom Fall, we realized that a very effective strategy for the enemy is to use the two techniques in concert. The enemy first deceives the expert system and then, just before and while commiting overtly to its true plan (e.g., surprise attack) defeats the system. Defeating the system masks the truth at the point when it would otherwise become apparent. The potential effectiveness of this strategy convinces us that "counter-defeat" deserves consideration alongside counter-deception.

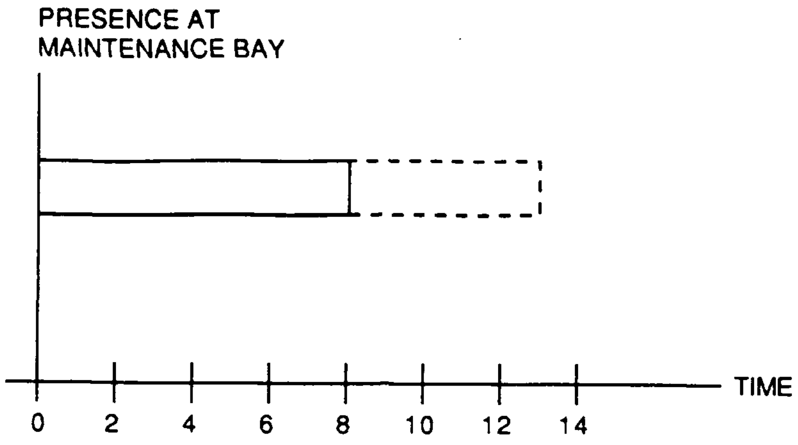

Figure 2-1: Apparent Behavior: Typhoon Maintenance

2.2 Sample Behavioral Analysis

The simplest way to motivate deception-behavior modeling and analysis is by means of an example. In Figure 2-1 we show an initial (friendly) behavior model for the maintenance of a Typhoon-class SSBN. This model just says that the minimum time for maintenance is 8 units and the maximum time is 13 units. This model is based on historical (satellite) observations. These models can be combined with observations to define the state of the SSBN over time and to hypothesize states for times when there are no observations. We depict this in Figure 2-2; we may assume that these are created by a suitable expert system. The observations use the notation "△" to indicate a positive siting of the presence of the SSBN, "▽" to indicate a negative siting, and "○" to indicate that observations are absent. The observations indicate a known presence for 9 units, which is consistent with the behavior model. The behavior model allows for up to 4 additional units. Since there are no observations on the front end a possible presence of 4 units is permitted. Since there is a negative observation on the back end the instantiated behavior is allowed just less than 3 units.

Now let us imagine that the analyst wonders whether more than one Typhoon could be maintained in this interval. The analyst might note that the 8 to 13 units of time apparently required to do a standard Typhoon maintenance imply Soviet SSBN maintenance efficiency is lower than other Soviet repair rates (e.g., MiG repair) and drastically lower than our rates. Thus the analyst sets out to explore the possibility that deceptions of various types

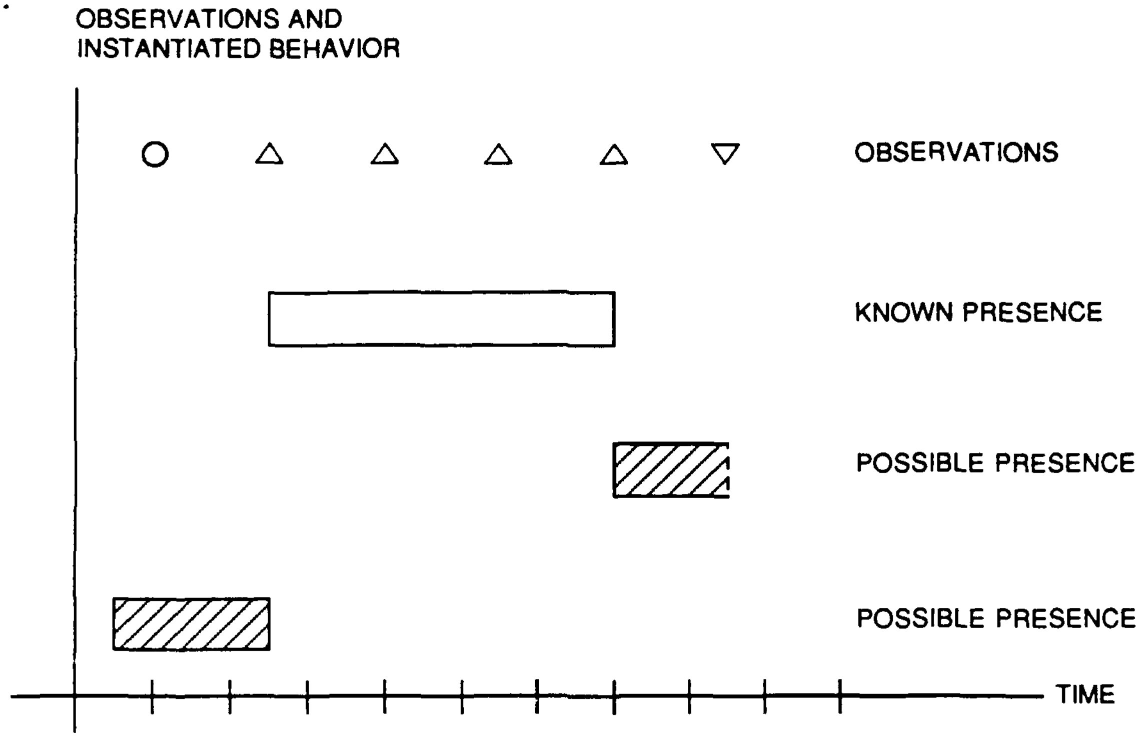

Figure 2-2: Instantiated Behavior Given Observations

might be aggregated so that two Typhoons get maintained but observations remain consistent with the behavior model stipulating 8 to 13 units.

The hypothesized behavior contains both deceptive behaviors (e.g., install IR decoy) and honest behaviors (e.g., maintain $Typhoon_2$). For convenience, we name a composite behavior that contains honest and deceptive sub-behaviors a "misleading behavior." The analyst's result is depicted in Figure 2-3. The basic schema here is that $Typhoon_1$ is maintained, then under the cover of an IR signature that mimics that of $Typhoon_1$, $Typhoon_2$ is maintained. Requirements for towing emerge in order to create maintenance space (i.e., physical and organizational constraints) and because exiting under power is denied (i.e., observation constraint). Loitering arises as a simple model that explains the behavior of $Typhoon_1$ when $Typhoon_2$ is being maintained; this is required because $Typhoon_1$ must physically persist (i.e., a physical or existence constraint) if it is to reappear later. Also, synchronization points are defined, such as that the decoy is turned off when $Typhoon_1$ exits (i.e., an observational constraint).

This misleading behavior is complex and cannot be created all at once; in particular, it is not filled out linearly in time. Instead, it is necessary to start with a goal (maintain two Typhoons), select deceptive behaviors, integrate these with honest behaviors and adjust constraints forward and backward in time, and derive the need for new behaviors or for assumptions.

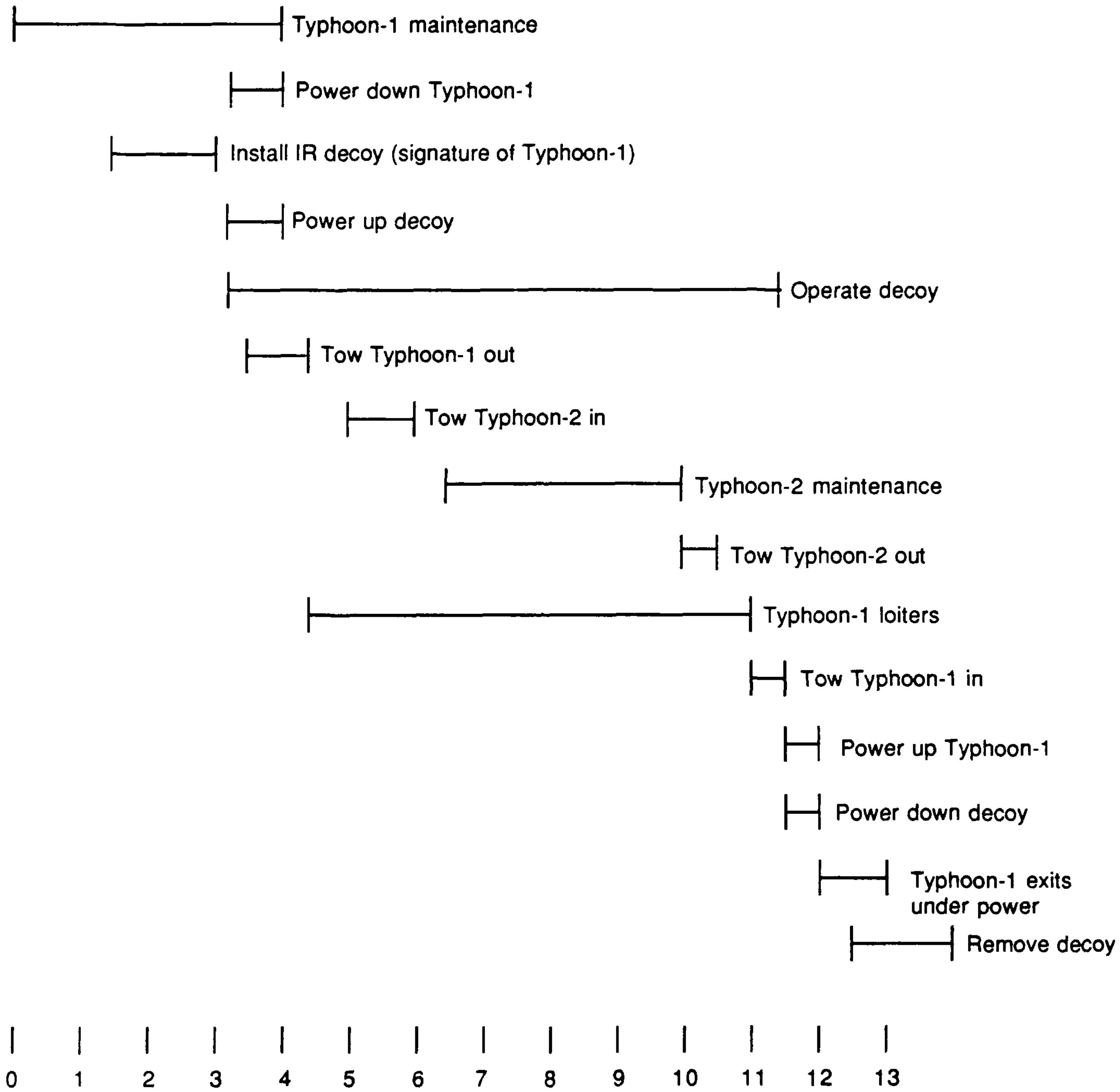

Figure 2-3: A Misleading Behavior: Maintaining 2 Typhoons

Once a misleading behavior is in place the analyst evaluates its characteristics. For example, is the deception realizable?[3] Is anything missing? Does the underwater topology near the maintenance bay support towing a Typhoon under the surface? Should any of our sensors have detected the towing operation? Depending on the answers to these questions, the behavior might be refined or even rejected.

If the behavior holds up under extensive testing then the analyst must investigate the implications for Soviet readiness and then for U.S./Soviet strategic parity. The analyst will probably also define evidence suggested by the deceptive elements of the behavior, that would tell him whether the behavior is indeed taking place. In the following section we discuss creation and analysis of a misleading behavior in technical terms.

2.3 Analyzing Behavioral Processes

We imagine the construction and analysis of misleading behaviors to proceed in the following way. Initially, the analyst posits some goals of the enemy (e.g., to double apparent maintenance). Then the analyst attempts to construct a feasible misleading behavior (or set of same) that satisfies the posited goal. This feasible behavior meets physical, temporal, and observational constraints for the internal sub-behaviors. Next, the analyst seeks to support or refute the misleading behavior(s) — we cast this as an argumentative process in which physics, economics, organizational behavior, military intentions, etc., are backgrounds in which the behavior is analyzed. During this stage of the analysis the analyst also reviews the created behavior to determine how friendly beliefs are being mislead and/or denied and also to establish the likely beliefs of the deceiver. A large part of belief analysis consists of analyzing the hypothesized enemy beliefs to see if they are consistent with other beliefs (e.g., the larger strategic intentions) attributed to the enemy. If the misleading behavior is feasible and if the beliefs that surround it are consistent with other beliefs of the enemy then the misleading behavior is analyzed for its impact on friendly/enemy parity (or disparity!).

We have made a possibly-subtle commitment here. The analyst starts with some beliefs of the enemy (the posited goals) but then defers reasoning about beliefs until a feasible set of behaviors has been produced that satisfies the goal. Only then does the analyst review the feasible misleading behavior to determine what its implications are for the beliefs of both parties. This approach has the effect of substantially narrowing the set of possible beliefs and

[3] We have thought of a simple, reliable, fairly inexpensive way to simulate the IR signature.

misleading beliefs that may be held by deceiver and deceived. To reason from beliefs alone does not appear to be even remotely tractable.

2.3.1 Construction of a Misleading Behavior

In our example we observed that several types of constraints attend the construction and execution of a misleading behavior. The most important ones seem to be physical, temporal, observational/perceptual, and organizational. Yet these constraints are not all stated 'up front;' in our example, we began only with the requirement to fit two hypothetical honest behaviors into the behavior shown in Figure 2-1. Thus we began with two maintenance behaviors (not necessarily of equal duration) that together require 13 units. We ended with shortened maintenance times and a broad collection of other supporting behaviors. For this report we content ourselves with describing the representation of behaviors and the basic strategy for constructing feasible misleading behaviors.

A *feasible misleading behavior* is a misleading behavior that satisfies certain posited goals and that satisfies physical, temporal, and observational constraints. Physical constraints include maintenance of dimension and extent, limitations on the number of artifacts and/or entities that can occupy a region of space, etc. Temporal constraints include sequencing requirements, minimal-residency requirements (e.g., this maintenance task requires this much time), etc. Observational constraints represent limitations on permissible emanation times and/or patterns.

A *behavior operator* models a behavior in the world and is represented as a state-transition operator along with pre- and post-conditions and a theory of what is happening while the operator is executing. Behavior operators are associated with artifacts (e.g., SSBNs) and entities (e.g., maintenance crews). The operator modifies the state of the artifact or entity: it accomplishes the basic action of the behavior. Pre-conditions are requirements to execute the operator, and post-conditions are things that are true of the artifact or entity after the operation. The theory attached to the state-transition operator represents things like electro-magnetic emanations and resource consumption that attend the operation.

For example, "`exit-maintenance-bay`" is a behavior-operator that contains a (state-transition) operator that changes the location of an SSBN from (the location of) a maintenance bay to (a location) outside of the bay. The pre-conditions are that the SSBN is in the bay, that the target location is empty of other SSBNs, and that a path exists from one location to the other. The post-conditions are revised readiness values for maintained systems, the bay is empty, etc. The attached theory describes heat emanations over time as a function of specific maintenance activities.

We conclude with one strategy for constructing misleading behaviors from behavior operators. The analyst constructs behavior operators and installs them as event-triggered modules in a design environment. Modules (behavior operators) can be triggered by the pre- or post-conditions left by other behavior operators. The environment starts with the posited goals and immediately derive a top-level behavior. For example, the goal of doubling Typhoon maintenance immediately leads to two back-to-back maintenance behaviors whose total extent is between 8 and 13 time units. Behavior operators are able to do two things: introduce new commitments and/or specialize existing ones left by other operators. Specialization occurs only over a range of variability left by prior operators; one operator cannot retract a commitment by another.[4] A misleading behavior exists when all pre-conditions are met and pre-conditions, post-conditions, and attendant theories form a coherent description of a behavior.

2.3.2 Analysis of a Misleading Behavior

We analyze a misleading behavior along evidential lines. That is, we view it as a conclusion and ask three questions: "What are its characteristics?", "What is its range of applicability?", and "What is its plausibility?". Detailed investigations will appear in our next report; for now we simply present the broad outlines.

Misleading behaviors introduce subsidiary processes and establish timing relations so that basic physical, observational, and organizational constraints are met. Initially, the way that it satisfies these constraints defines its main properties. But other properties can be defined, such as the sensitivity of the deception to special conditions, its general reliability (e.g., measured as a function of synchronization requirements), etc. Economic, organizational, and political relationships may be definable for the behavior. Ultimately, the analyst is creating a larger model that shows how this misleading behavior fits in to the overall scheme of enemy capabilities and actions. This stage of analysis also suggests tests that an analyst might make (e.g., new observations that would support or refute the misleading behavior).

Analysts look at the range of application of a deception to determine where and when it will be used. A deception that can be used in a broad range of circumstances is (in general) more valuable than one that cannot.

[4] This is handled by allowing operators to spawn new environments in which their commitment, and not the conflicting one, is made.

The plausibility of the deception is based on the foregoing analyses. We treat the behavior as feasible if it passes physical, temporal, and observational constraints (i.e., these are hard constraints). For feasible behaviors we define plausibility according to the degree of difficulty it requires to create and maintain it. This degree of difficulty is defined along the following dimensions (at least): economic, organizational, political, transportation, resource, and deeper models of maintenance operations. These dimensions can be viewed as soft constraints. Finally, the plausibility is adjusted according to its range of applicability, under the assumptions that a wider range is correlated with greater interest or stronger intentions on the part of the enemy to use it — leading to greater plausibility for the construction and use of the behavior.

3. Future Investigations

3.1 Research Issues

The following issues will be considered as we continue to study the problem of deception and the needs of counter-deception.

- Principles that underwrite design of expert systems, such as:
  - encode an explicit concept of deception, be it within the expert system or an augmentation of the expert system,
  - model the enemy as an intelligent actor motivated to deceive (i.e., encode deceptive intentions and behaviors),
  - explore contradictory indications (i.e., explore the possibility that a contradiction is a result of a breakdown in a deceptive behavior, and that by embedding a failed deceptive behavior within an honest behavior the contradiction can be explained), and
  - encode models of the weaknesses of sensors (unobserved mechanics of deception) and strengths (opportunities for cross-validation of evidence),
- Organizational support, such as:
  - train analysts in deception techniques,
  - consider the possibility of selecting analysts in part for natural ability to penetrate deception.
- Basic support techniques/mechanisms the system should have:
  - tools to explore sensor system gaps (that can be exploited) and also tools to suggest and/or plan for the use of several sensors to cross-validate observations,
  - tools to create and/or modify internal components to explore the effect deception models would have on conclusions (i.e., modify the expert system to see if it is being fooled),
  - given sensor and self models, tools to explore the value of gathering evidence,
  - evidential tools (reasons and degree of belief) that are sensitive to the effect of deception, and

- reflective capabilities to reason about alternative constructions.

- Expert-system environment tools to assist the analyst, such as:

  - ability to partition information and construct conclusions from each partition.

  - tools to gauge responses, based on the product of the expert system, and possibly de-sensitize them to the possibility of undetected deception (i.e., reduce value of enemy deception).

3.2 Sample Domains for Counter-Deception

We require sample domains to test our ideas. These domains must be such that expert systems exist or can be constructed to assist in the domain, and in which credible deceptive behaviors can be imagined. Our approach will be to describe the sample exper* system and its main source(s) of evidence, and then determine what it would take to deceive the system. This becomes the background for our counter-deception concepts and tools. We believe that it will be useful to examine more than one domain since we are interested in a theory of counter-deception that can be broadly applied.

We have thought of three domains for counter-deception. We list them, along with their main source of evidence:

- Readiness of Soviet SSBNs: we analyze the apparent versus true maintenance state, along the lines of our examples in the previous chapter, and look at the impact on apparent versus true readiness.

  The main sources of evidence are satellite data, information on maintenance across various Soviet military commands, and physical specifications of the local environment. The last includes surrounding terrain and cities, road and rail networks, and underwater topology.

- Soviet naval strategy: analyze their espoused strategies to see if it is consistent with what they appear to be building and doing.

  The main sources of evidence include a specification of their existing fleet (and ours), analyses of the relation between strategy and hardware, production plans, and our espoused strategy.[1]

[1] The book by Breemer, mentioned in Appendix B, contains useful source material on the relation between hardware acquisition and underlying strategy for employing submarines. In particular, he discusses the suitability for first and second strike roles based on technological capabilities and acquisition considerations. He also discusses the issue of whether Soviet writers on naval strategy are publishing statements of policy or instead are arguing for a new policy; clearly, the publication of such a document could also be meant to deceive.

- Political/organizational ascendancy: who are the primary individuals and what are their relationships? Who are the true decision-makers and who are the figureheads and bureaucrats? The expert system may be looking for policy-makers in specific agencies; it could equally well be tracking terrorists or drug smugglers.

  Published reports, newspapers, news commentary, and police or intelligence reports provide the evidence for this expert system. (Expert systems of this sort exist and could be easily deceived.)

We remark that the first domain has several very desirable aspects:

- The strategic consequences of deception in this domain are profound.
- ADS has very deep knowledge of this problem, plus we are aware of expert systems that address precisely this problem.
- This is a long-term problem. In general, this means we have a lot of historical data that we can analyze. Also, it has the important property that deception in this domain is not a one-shot event — giving us a much better chance of detecting them.

  As a long-term deception it must be responsive to improvements in our sensing capabilities. We can examine how improvements in our observation facilities constrain deceptive behaviors. It also gives us a way to think about anomalies, since we expect that they must make mistakes in a long-term deception.

3.3 Measurement of Our Work

A key issue for us is the measurement of our work. We will not consider our work to have been successful if all we do is provide a collection of tools that allow an analyst to create and analyze complex feasible behaviors that contain deception. After all, they may have no bearing on extant deceptions. Thus, it is necessary to find a way to test the support tools.

We have thought of two approaches. One is to reconstruct past deceptions (e.g., in World War II), along with a suitable expert system, and determine whether our concepts and tools can penetrate the deception. The second is to conduct a sequence of experiments with an expert system and two teams. "Red team" attempts to deceive the expert system. "Blue team" has the expert system, along with our tools, and attempts to pentrate the deception.

The experiments are parametrized by sensor capabilities and for varying degrees of information that Red team has about the expert system.

Of course, one other approach would test our system. If the government were to gather (a new type of) data as a result of a conjecture regarding deception, and if this new data exposed a deceptive behavior, then we would view it as a success!

A. Logical Belief Relations Under Deception

We need a language to represent the beliefs of agents. This language must, as a special requirement, also express beliefs about the beliefs of other agents. We will be using this language to expose special relationships in beliefs that arise under deception.

"Belief" is an abstraction that we use to describe and refer to a cognitive item of an agent. We do not introduce a particular cognitive apparatus here, so we immediately restrict "belief" to mean a sentence that is held by an agent. Specifically, beliefs are only associated with agents and they are represented as a relation over agent names and sentences in a language. We denote this $Bel(\alpha, S)$, where α names an agent, S is a sentence in a suitable language, and Bel names the belief relation. For example, $Bel(Analyst, (readiness\ NorthernFleet\ 0.35))$ represents the belief of an agent named "Analyst" that the readiness of the northern fleet is 0.35. The language must allow sentences that are beliefs: $Bel(Agent_1, Bel(Agent_2, T))$. This is necessary to represent beliefs about the beliefs of other agents.

Construction of the language is a delicate operation. One approach is to start by defining an "internal language" $\mathcal{L}$, such as the well-formed formulas of first order predicate calculus. $\mathcal{L}$ is then the target of an outer, modal language $\mathcal{M}$: beliefs are sentences that attribute a sentence in $\mathcal{L}$ to an agent via the modal operator *Bel*. *Bel* is a modal operator since one of its arguments is a sentence in the internal language; the construct $Bel(agent, S)$ is called a modal construct. Formation rules for modal constructs also permit nesting of modal constructs: the sentence S is allowed to be a sentence in $\mathcal{M}$, not just in $\mathcal{L}$. Thus the following belief expresses a belief of "EnemyAgent" about a belief of "FriendlyAgent:" $Bel(EnemyAgent, Bel(FriendlyAgent, (readiness\ NorthernFleet\ 0.35)))$.

In general, languages (e.g., $\mathcal{L}$ and $\mathcal{M}$) are characterized by their expressiveness and are sanctioned according to whether they support sound inference relative to the formation rules and a set of deductive schemata that may be applied to the sentences of the language. Languages to represent the beliefs of other agents have been invented not only to represent beliefs of (other) agents but also to represent reasoning of others. For example, suppose that $Agent_1$ believes that $Agent_2$ believes both P and $P \rightarrow Q$. Also, suppose $Agent_1$ believes $Agent_2$ believes in the deduction rule called "Modus Ponens," namely: $(P \text{ and } P \rightarrow Q) \mapsto Q$.[1] Finally, suppose that $Agent_1$ believes that

[1] The "$\mapsto$" symbol is used to display the antecedent and result of a mapping.

$Agent_2$ constructs conclusions that follow from its beliefs and that it has had time to do so. ***Then*** $Agent_1$ ***can deduce that*** $Agent_2$ ***believes*** Q.

How does deception enter? Deception appears as systematic differences between what an agent believes about itself versus what it believes others believe about it. These differences presumably are caused by deceptive behaviors. For example, the Soviets have a high payoff in deceiving us such that we believe that their readiness is lower than it actually is. This can be represented as the pair of beliefs:

- $Bel(EnemyAgent, Bel(FriendlyAgent, (readiness\ NorthernFleet\ 0.35)))$
- $Bel(EnemyAgent, (readiness\ NorthernFleet\ 0.55))$

To emphasize that this difference is a result of deception, let us extend our belief notation to include justifications. Let "[X; Y]" denote that the belief X is justified by the belief set Y. Also, let us simplify our belief notation by suppressing the modal relation. Let "E :" stand for "enemy believes" and "F :" stand for "friendly believes." We introduce "ES_F" to represent the model of the friendly expert system. "ES" is a straightforward expert system that processes observations under the assumption that all behaviors are honest. Then the beliefs for a perfect deception could be represented as the following system:

- $E : [S_E = \Sigma; B \mapsto \Sigma]$
- $E : [ES_F = ES; deception\ is\ impenetrable]$
- $E : [F : S_E = T; E : O_F = \Theta, \Theta(B) = \epsilon]$
- $F : [S_E = T; \Theta(B) = \epsilon, ES(\epsilon) = T]$
- $E : [\Sigma \succ T; NetValue(\Sigma) > Value(T)]$

Here S_E stands for the state of the enemy, B stands for the behaviors (honest and deceptive) executed by the enemy, and T is the friendly view of the enemy state, based on its observation model O_F. The value of O_F for friendly is "Θ"; the enemy has correctly modeled this. Application of Θ to the behaviors B yields observational evidence "ϵ." "$\succ$" stands for a preference ranking.

The first belief represents the enemy's (correct) model of its true state, based on the known results of executing the behaviors B. The second belief correctly asserts that the friendly reasoning process is modeled by "ES." The combination of the third and the fourth beliefs expresses that the deception has worked: the enemy belief that friendly has the desired false belief is in fact the friendly belief. It also represents that the enemy has the correct model of the friendly observation process. The fifth belief expresses that the true enemy state, Σ, is preferable (for the enemy) to T since the value of Σ, accounting for the cost of the deception, exceeds the value of T.

We can also represent the configuration of beliefs that arises when a deception has failed. We now introduce "ES^+" — this is "ES" augmented with counter-deception support; it uses a wider field of evidence which is supplied by the augmented observation process "Θ^+". The configuration under total failure of the deception (i.e., complete friendly penetration of the enemy deception) is:

- $E: [S_E = \Sigma; B \mapsto \Sigma]$
- $E: [ES_F = ES; deception\ is\ impenetrable]$
- $F: [ES_F = ES^+, O_F = \Theta^+; deception\ is\ analyzable]$
- $E: [F: S_E = T; E: O_F = \Theta, \Theta(B) = \epsilon]$
- $E: [\Sigma \succ T; NetValue(\Sigma) > Value(T)]$
- $F: [S_E = \Sigma; \Theta^+(B) = (\epsilon, \kappa), ES^+(\epsilon, \kappa) = \Sigma]$
- $F: [E: F: S_E = T; F: E: F: ES_F = ES, ES(\epsilon) = T]$

The new beliefs express the result of augmenting the observation process and the expert system. The last belief represents the friendly model of the way the enemy models the friendly reasoning process.

The lesson here is that the language of beliefs can depict in a precise manner the belief relations that arise under deception. It is possible to study these relations abstractly to a limited extent — several different types of belief relations can be derived that characterize different types of deceptions and (especially) the success or failure of the deception.

However, it should be clear from the main text and also from this example that the heart of the analysis of deception is the problem of orchestrating processes to deceive. Our current conclusion is that the language of belief relations can depict the results but gives us little guidance when we seek to combine deceptive and honest behaviors. Equivalently, we believe that it is feasible to use feasible deceptions as a foundation for speculating about underlying beliefs of the enemy, but that the reverse process will never be solved. The reason is clear once we recognize the combinatoric realities. The possible beliefs of the enemy are tremendous, and even the hypothesis of enemy beliefs does little to restrict the feasible deceptive behaviors. Hence reasoning from beliefs to deceptions holds little promise. But the possible deceptions are limited by physical, temporal, and observational factors, and once feasible misleading behaviors have been constructed the realm of feasible enemy beliefs (consistent with the misleading behaviors) has been significantly constrained. Hence reasoning from feasible deceptions to beliefs may be tractable.

B. References

Breemer. *Soviet Submarines: Design, Development, and Tactics.* Jane's Information Group, UK, 1989.

Brown. *Bodyguard of Lies.* Bantam Books, 1976.

Cruikshank. *Deception in World War Two.* Oxford University Press, 1979.

Daniel, Herbig. *Strategic Military Deception.* Pergamon Press, 1982.

Greenberg. The Role of Deception in Decision Theory. In *Journal of Conflict Resolution.*

Greenberg. The Effect of Deception on Optimal Decisions. In *Operations Research Letters.*

Heuer. Strategic Deception and Counterdeception. In *International Studies Quarterly*, vol 25, no 2, 1981.

Mihalka. Soviet Strategic Deception. In *Journal of Strategic Studies*, Vol 5, no 1, 1982.

Whaley. Toward a General Theory of Deception. In *Journal of Strategic Studies*, March, 1982.