MTR140412
MITRE TECHNICAL REPORT

MITRE

Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking (ICArUS):

Phase 2 Challenge Problem Design and Test Specification

Kevin Burns

November, 2014

This publication is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) ICArUS program, BAA number IARPA-BAA-10-04, via contract 2009-0917826-016, and is subject to the Rights in Data-General Clause 52.227-14, Alt. IV (DEC 2007). Any views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Approved for Public Release; Distribution Unlimited 14-3960

McLean, VA

Abstract

Phase 2 of the IARPA program ICArUS (Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking) requires a research problem that poses cognitive challenges of spatial-temporal sensemaking (BAA, 2010). The problem serves as a challenge for performers who are building integrated cognitive-neuroscience models, and as a tool for obtaining data from human experiments. This document describes the challenge problem, and outlines the T&E (Test & Evaluation) approach for evaluating models in Comparative Performance Assessment and Cognitive Fidelity Assessment (BAA, 2010). Normative (Bayesian) solutions to the challenge problem are derived, as needed to support the assessment of human and model performance. Opportunities are also identified for transition of the challenge problem design and results to the geospatial Intelligence Community.

Note: This document was originally prepared and delivered to IARPA in March, 2014, in order to support ICArUS Phase 2 T&E efforts that concluded in June, 2014.

Table of Contents

List of Figures

List of Tables

1 Introduction

This document was originally prepared and delivered to IARPA in March, 2014, to support ICArUS Phase 2 Test & Evaluation (T&E) efforts that concluded in June, 2014. Further background is provided in a summary document (Burns, Fine, Bonaceto, & Oertel, 2014) titled *ICArUS: Overview of Test and Evaluation Materials*, available at http://www.mitre.org/publications.

The ICArUS Phase 2 challenge problem is a ***Tractable Analytic Challenge To Investigate Cognitive Sensemaking***, dubbed TACTICS. The design is a balance of experimental rigor, for assessment of models in the laboratory, and practical relevance, for transition of results to real-world applications in the Intelligence Community. This balance is achieved using a computational approach to human experiments and model evaluations, covering a spectrum of "missions" that are all *Variations* (Section 3) on the same basic task *Description* (Section 2). Normative *Solutions* (Section 4), which are needed for rigorous *Evaluation* (Section 5), are developed as part of the design. Important *Definitions* (Section 7) and a brief discussion of potential directions for long-term *Transition* (Section 6) are also provided.

Referring to the title of this document, Sections 1-7 all address the challenge problem "design". The "test specification" is captured in Section 5 (*Evaluation*), which describes the methods and metrics for various assessments required by the BAA (2010).

1.1 Motivation

Although practical applications to real-world intelligence are not the focus of this document, TACTICS is intended to aid ***Transition And Communication To Intelligence Community Stakeholders***. This objective is accomplished using a computational approach to human-experimental design and a relational mapping to real-world intelligence analysis.

The relational mapping to support *Transition* (Section 6) is based on computational variables made explicit in the design of TACTICS. More specifically, six types of intelligence analyses (and corresponding variables of TACTICS) are characterized as: *vulnerability* analysis (P), *opportunity* analysis (U), *capability* analysis (P_c), *activity* analysis (P_t), *frequency* analysis (F_t), and *intentionality* analysis (P_a). TACTICS addresses all six, but focuses on how these various analyses are integrated in ***sensemaking***. The six types of analyses and corresponding variables of TACTICS are explicitly mapped to 26 real-world case studies of geospatial intelligence. These case studies were developed in Descriptive (Cognitive) Task Analysis (MITRE, 2013), via interviews with analysts and reviews of published articles, see *Transition* (Section 6). As noted in *Transition* (Section 6), TACTICS is:

> *A game of repeated risk assessment and action (Kaplan & Garrick, 1980; Garrick, et al., 2004), posing cognitive challenges that are prototypical of intelligence and operations in threat situations (Burns, 2010; McDonald, 1950) – including counterinsurgency (COIN) and other security domains (airport/border, cyber/network, crime/fraud, drugs/gangs, etc.).*

Here the term "game" (von Neumann & Morgenstern, 1944) is used in the game-theoretic sense of an adversarial (Red-Blue) interaction requiring inferencing and decision making – including inferences about *what*, *when*, and *where* the opponent will act (an ***action***), *how* he will act (a ***tactic***), and *why* he will act that way (an ***intent***).

1.2 Foundation

With respect to rigor, a computational approach to challenge problem design begins by formalizing *Definitions* (Section 7) of conceptual notions described in the BAA (2010), especially the notion of a "frame" and associated "core sensemaking processes" listed in Table 1 of the BAA. Here at the outset it is useful to highlight a few of these definitions, first and foremost that of *sensemaking* (where italicized words are all defined in Section 7):

> ***Sensemaking*** is a recurring cycle of obtaining *evidence* and updating *confidence* in competing *hypotheses*, to *explain* and *predict* an evolving situation.

This definition is consistent with literature cited in the BAA, including Klein, et al. (2007), who cite Weick (1995), who cites Louis (1980), who named and described the process as follows:

> *"Sensemaking can be viewed as a recurring cycle... The cycle begins as individuals form unconscious and conscious anticipations and assumptions, which serve as predictions about future events. Subsequently, individuals experience events that may be discrepant from predictions. Discrepant events, or surprises, trigger a need for explanation, or post-diction, and correspondingly, for a process through which interpretations of discrepancies are developed..."*

According to this description, sensemaking can be boiled down to three basic processes by which humans "make sense" of any real-world situation (Burns, 2014; 2005) or media communication (Burns, in press; 2012), as follows: First a person uses current beliefs (confidences in hypotheses) to form ***expectations*** of data (evidence). These expectations may or may not be met by subsequent observations. Any ***violation*** of expectation, from surprising evidence, then fuels the formation of an ***explanation*** – which is an updating of beliefs (confidences in hypotheses) in light of the data (evidence).

Moving beyond this conceptual description, a comprehensive understanding of sensemaking requires computational modeling at functional, psychological, and biological levels. Although the latter levels are the main aim of ICArUS, design of a challenge problem first requires a computational theory at the functional level, in the Marr (1982) sense of specifying *"what is the goal of the computation..., and what is the logic of the strategy by which it can be carried out?"*

One such theory (dubbed Octaloop; see Burns, 2014) was developed to guide design of the Phase 1 challenge problem (Burns, Greenwald, & Fine, 2014), and the same theory is used here to guide design of the Phase 2 challenge problem. By necessity, this computational theory goes further than conceptual notions like those of the "data-frame" theory (Klein, et al., 2007) described in the BAA. In particular, the term "frame" is used loosely by many authors (cited in Klein, et al., 2007) to mean many different things. The data-frame theory itself never defines

"frame" precisely, but rather uses this term in referring to a "story", "map", "script", "plan", or any other explanatory knowledge structure that is not data and yet is needed to make sense of data. Here the term is given a more formal definition as follows:

> ***Frames*** are knowledge structures, comprising *hypotheses*, *evidence*, and *confidences*, including conditional *likelihoods* of *evidence* (i.e., conditional on *hypotheses*) as well as conditional *likelihoods* of *hypotheses* (i.e., conditional on *evidence*). In ***spatial context frames***, *likelihoods* depend on spatial factors. In ***event sequence frames***, *likelihoods* depend on temporal (and spatial) factors.

When the components of frames are made explicit, as in this definition, researchers are in a better position to model and measure how frames might be "learned" and "assessed" and "re-framed" – as all of these terms are used to describe "core sensemaking processes" in BAA Table 1. In particular, the notion of *re-framing* is defined more formally here as follows:

> ***Re-framing*** **(aka** ***Set-shifting*****)** is a revision of *hypotheses*, or revision of *confidences* across *hypotheses*, in which the most likely *hypothesis* changes due to the observation of surprising *evidence* (i.e., *evidence* that is not likely to be caused by the currently-most-likely *hypothesis* or *hypotheses*).

Besides distinguishing between *hypotheses* and *confidence*, the computational definitions above also distinguish between *hypotheses* and *evidence*. This difference is important because it reflects ***causal structure*** (Pearl, 2000), which plays a key role in all sensemaking – including *forward (prognostic) inferences* whereby a sensemaker is forming *expectations* – as well as *backward (forensic) inferences* whereby a sensemaker is forming *explanations*. Thus the causal structure is ***hypotheses → evidence***, where hypotheses are hypothetical causes of evidential effects (i.e., causes → effects) and the direction of inferencing can be in either or both directions – forward along the arrow direction or backward in reverse of that direction. A ***causal hierarchy*** is merely the nesting of this basic structure into more complex structures where hypotheses at one level serve as evidence at the higher levels (see Figure 3 of Burns, 2005).

TACTICS is based on a causal hierarchy with four arrows as follows:

> **intent → tactic → action → feature → datum.**

The task itself requires re-framing at each level of the causal hierarchy, as discussed further in Section 1.3 (*Clarification*). Mathematically, causality at each level is measured and modeled by conditional probabilities – and these conditional probabilities are computational representations of ***event sequence (and spatial context) frames***. Conceptually, the five levels in this causal hierarchy are similar to the Joint Directors of Laboratories (JDL) Data Fusion Group model. The JDL model (Steinberg & Bowman, 2004) is a functional-hierarchical specification of input data, model outputs, and associated *inferencing* applicable to a broad class of geospatial fusion problems aimed at understanding and affecting situations (similar to sensemaking, but with a focus on system performance rather than human performance). The five layers of the JDL model, labeled 0 (Raw Signals), 1 (Entities), 2 (Situations), 3 (Impact), and 4 (Performance), can be mapped roughly to the TACTICS levels of *datum*, *feature*, *action*, *tactic*, and *intent*, respectively.

1.3 Clarification

Per BAA Table 3, Phase 2 of ICArUS is focused on a notion of event "sequences", and associated cognitive biases that may arise from heuristic processes in human sensemaking. The purpose of the present section is to clarify how TACTICS captures sequences, and how this treatment relates to previous literature on "frames" (noted above) – especially "scripts".

Temporal events in the form of "sequences" are often referred to as "schema" (Barlett, 1932) or "scripts" (Schank & Abelson, 1977). For example, one sequence may be A, B, and C, where B is likely to occur after A, and C is likely to occur after A and B. Such scripts (or plans or event sequence frames) are formally defined by conditional probabilities, e.g., P(B|A) is high and P(C|A,B) is high. Importantly, it is only through knowledge of these conditional likelihoods that a sensemaker can make *predictions* like "probably C next" after observing A and B; also form *explanations* like "probably script 1" after observing all or part of the sequence A, B, and C.

In TACTICS these sorts of scripts occur at three different time scales in nested levels of the causal hierarchy. At the lowest level (and shortest timescale), a player receives a sequence of intelligence reports (aka INTS), each reporting some ***datum***. From these data the player infers temporal-spatial ***features*** that relate to different stages of an ***action*** script – e.g., the enemy *vulnerability* (a spatial feature), *capability* (a temporal feature), and *activity* (a temporal-spatial feature). This sequence is akin to a sequence A, B, and C described above, where the analogue of "script 1" is "attack" and "script 2" is "no attack".

Then, at a higher level of the causal hierarchy (and longer timescale), the sequence is a series of actions such as "attack", "no attack", "no attack", etc. Once again the sequence is governed by conditional probabilities that depend on spatial and temporal context. In this case the scripts lie at the level of ***tactics***, e.g., "tactic 1" and "tactic 2", where an enemy who plays with tactic 1 (e.g., aggressive) is likely to exhibit a different pattern of actions (attacks) than an enemy who plays with tactic 2 (e.g., passive). Knowledge of these tactics, including their underlying conditional probabilities, is what enables a player to predict actions (attack or no attack) from assumed tactics, and also to infer tactics (tactic 1 or tactic 2) from attack patterns.

Finally, at an even higher level of the causal hierarchy (and even longer timescale), a script is a sequence of tactics such as "tactic 1", "tactic 2", etc., where a Blue player must explain and predict changes in Red tactics that are governed by enemy ***intent***.

Notice that the notion of set-shifting applies at each of the three levels and timescales described above. For example, at the highest level a player may know or learn that his opponent is consistently playing according to tactic 1 (e.g., aggressive). So "tactic 1" becomes a strong assumption and the player is led down a so-called "garden path" of expectations. The set-shift then comes after a surprise (Burns, in press; 2012), when the player is faced with overwhelming evidence to the contrary. This is a ***violation*** of ***expectations***, which requires re-framing in order to form an ***explanation*** like ***"Aha – tactic 2!"***.

Likewise, set-shifting happens at a lower level when a player strongly expects an attack and is surprised to observe no attack (or vice versa). This forces re-framing of beliefs about how actions

are constrained by tactics (and intents). Finally, set-shifting also occurs at an even lower level when the player strongly expects one feature from INT data and yet observes a different feature. This forces re-framing of beliefs about how spatial-temporal features of INTS are constrained by intentional actions.

As described above, set-shifting in TACTICS differs in three important ways from other laboratory tasks more typically used for measuring the phenomenon, such as the Wisconsin Card Sorting Task (Berg, 1948; Monchi, et al., 2001). One difference is that in TACTICS the so-called "rule" (or "script") is not deterministic but rather it is probabilistic, governed by conditional probabilities. The reason for this is that a probabilistic task is required to capture the relevant conditions of real-world situations in which set-shifting (and sensemaking more generally) actually occurs, i.e., under uncertainty. A second difference is that in TACTICS the set-shifting occurs at three different (nested) time scales, namely: within a trial (*feature* set-shifting); between trials (*action* set-shifting); and between batches of trials (*tactic* set-shifting).

A final difference is that in TACTICS the set-shifting occurs in a *causal* hierarchy, at each level of the hierarchy as well as across levels of the hierarchy. Moreover, and perhaps most importantly, *intent* is itself constrained at the highest level of the hierarchy via a reward structure given by the payoff matrix of the game (see *Description*, Section 2). The reward structure provides players with a natural basis for causal reasoning, as it encourages and enables them to explain *why* there was a change – not just *how* things may have changed or *what* (or *when* or *where*) things may have changed. This feature of a game allows the laboratory task to more realistically capture the causal structure of naturalistic situations that are relevant to real-world intelligence analysis and security operations (Burns, 2010).

1.4 Missions

Besides the *inferencing* processes that are central to re-framing (set-shifting) across a causal hierarchy, as discussed above, TACTICS poses additional cognitive challenges that are associated with many real-world sensemaking situations. These processes, which are addressed in *Variations* (Section 3) of the basic task, include *decision-making* based on inferences and *foraging* for new evidence. TACTICS addresses all three cognitive processes, i.e., ***inferencing***, ***decision-making***, and ***foraging***, in order to cover the scope of sensemaking set forth in the BAA as follows:

> *"Sensemaking is a volitional process that involves multiple shifts in attention, continuous exploration [**foraging**], and evaluation [**inferencing**] of multiple pieces of evidence, and repeated **decision making**..."*

The design of TACTICS includes various "missions" that address each of these three processes, individually (to the extent they can be separated) and in combination. But before discussing *Variations* (Section 3), the basic task is presented first in *Description* (Section 2).

As an overview, Figures 1 and 2 are screen shots of the graphical user interface use in the missions. Many more screen shots and non-technical instructions to users are provided by the tutorial (see Burns & Bonaceto, 2014) embedded in the TACTICS software itself.

Table 1 provides a listing of variables referred to in the *Description* (Section 2) and *Variations* (Section 3). Table 2 summarizes the temporal-spatial *features* of intelligence *data* (sources) modeled in TACTICS, along with the *meaning*, *measure*, and *symbol* assigned to each feature.

As seen in Table 2, the measure of meaning for each feature is a probability (frequency), which is a measure of likelihood; or a utility, which is a measure of consequence. This is because TACTICS involves *inferencing*, which is computationally modeled by ***probabilities***; as well as *decision-making* and *foraging*, which are computationally modeled by expected ***utilities***. As discussed in Section 6 (*Transition*), raw data are of no use in accomplishing these cognitive competencies unless some person or system infers or assigns associated likelihoods (probabilities) and consequences (utilities).

In TACTICS, most of the probabilities and utilities are assigned to raw data by INT sources themselves – much like real-world intelligence would provide some *measure* of *meaning* beyond just raw *data*. This is to focus ICArUS experiments on the cognitive processes of sensemaking per se, rather than on estimating various quantities needed as input to sensemaking. The approach also enables experimental measures of "average" sensemaking performance (as required by BAA), where the average is an average over human subjects who are all using the same inputs to sensemaking.

The main exception to this approach involves a Blue intelligence handbook called the BLUEBOOK, which represents Red tactics as needed for Blue to infer the propensity (likelihood) of Red attack. In some cases, Red tactics are not known for sure and hence must be inferred forensically from past attacks (SIGACTS). For those cases, the input to prognostic sensemaking involves a good deal of forensic sensemaking, i.e., in a mission where Blue must infer Red tactics and detect changes in Red tactics (see *Variation*s Section 3).

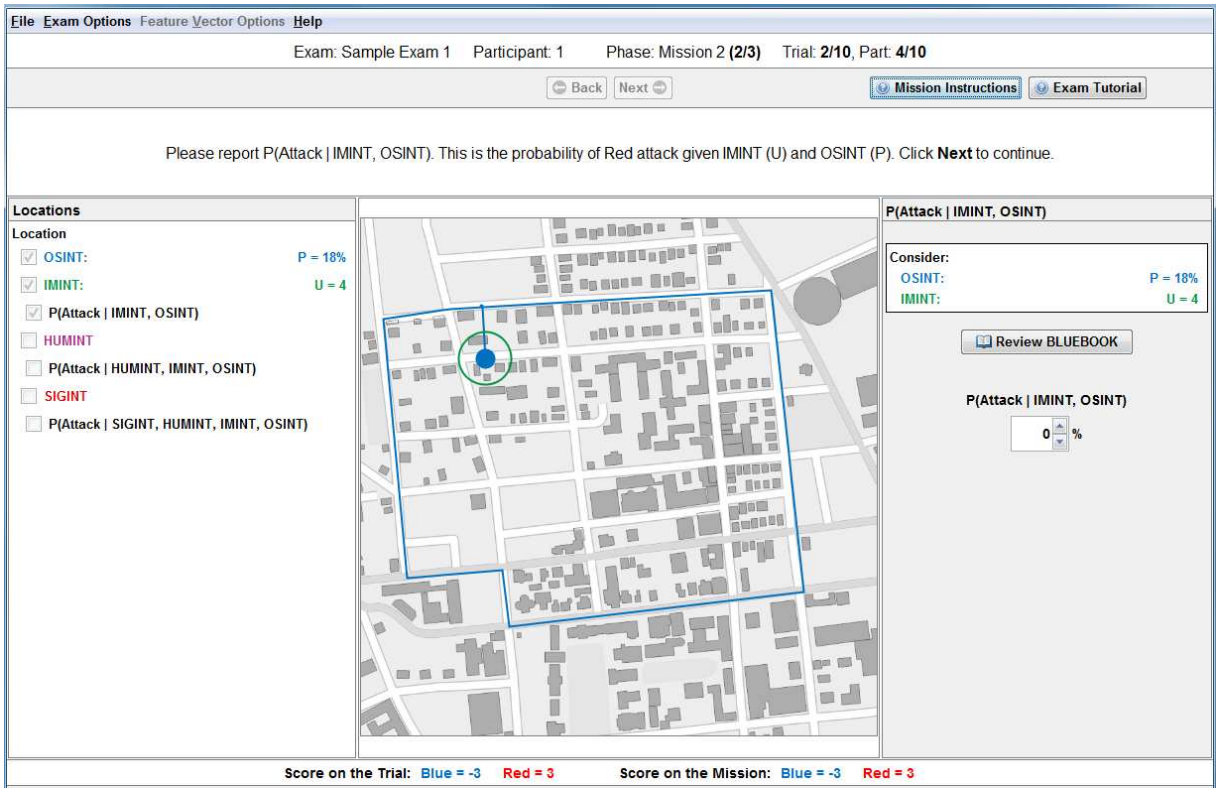

Figure 1: Geographic Information System (GIS) display and Graphical User Interface (GUI).

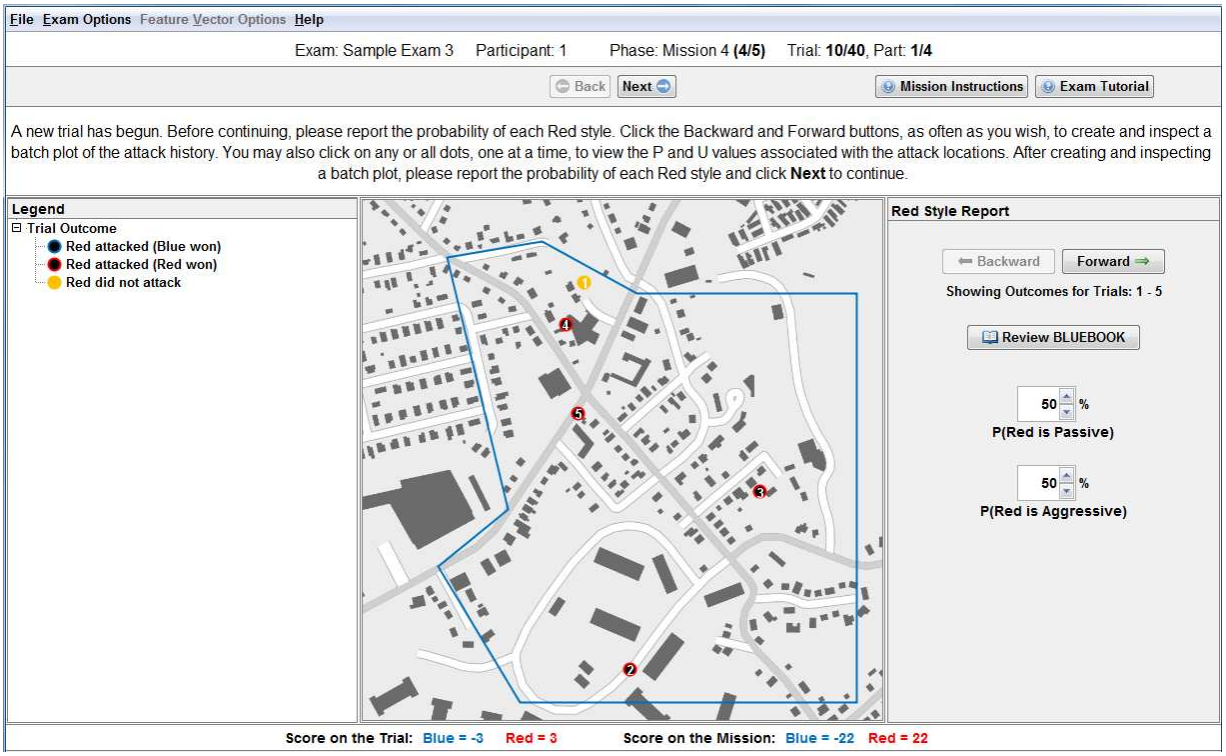

Figure 2: "Batch plot" of significant activities (SIGACTS) for a series (batch) of trials.

Table 1: Listing of variables in design of TACTICS.

Symbol	Meaning
a	attack, an action by Red
~a	not-attack, an action by Red
B_B	Blue's model of his own (Blue) tactics
B_R	Blue model of his opponent's (Red's) tactics
B_t	Blue's choice of action (d or ~d) on trial t
d	divert, an action by Blue
~d	not-divert, an action by Blue
F_t	frequency of past activity by Red over some number of trials (t)
P	probability that Blue will defeat Red in a showdown (i.e., if a and ~d) at a Blue point, P(x,y)
P_a	probability that Red will attack on trial t, $P_a(t) = P_{t,p,c}(t)$
P_c	probability that Red has the capability to attack on trial t, $P_c(t)$
$P_{\sim d}$	probability that Blue will not divert on trial t, $P_{\sim d}(t)$
P_p	probability that Red has the propensity to attack on trial t, given the capability to attack, $P_p(t)$
$P_{p,c}$	probability that Red has the propensity and capability to attack on trial t, $P_{p,c}(t)$
P_t	probability of Red attack as signaled by Red activity on trial t, $P_t(t)$
$P_{t,p,c}$	probability of Red attack on trial t, per activity, propensity, and capability, $P_a(t) = P_{t,p,c}(t)$
r	shortest straight-line distance from Blue point to Blue border
R_B	Red's model of his opponent's (Blue's) tactics
R_R	Red's model of his own (Red) tactics
R_t	Red's choice of action (a or ~a) on trial t
t	trial number; also number of trials in F_t or number of trials since last attack in function for $P_c(t)$
U	utility at stake in a showdown at a Blue point, U(x,y)
v	constant parameter in vulnerability function for P(x,y)
x,y	space coordinates

Table 2: Temporal-spatial features of intelligence data in TACTICS, along with the associated meanings, measures, and symbols.

Datum	**Feature**	**Meaning**	**Measure**	**Symbol**
OSINT	Proximity	Vulnerability	Probability	P
IMINT	Density	Opportunity	Utility	U
HUMINT	Recency	Capability	Probability	P_c
SIGINT	Reliability	Activity (prognostic)	Probability	P_t
BLUEBOOK	Probability and Utility	Propensity	Probability	P_p
Batch Plots (SIGACTS)	History	Activity (forensic)	Frequency	F_t

2 Description

Phase 2: The text of this *Description* will focus on the basic task, to be implemented in five "missions" of the Phase 2 experiment (see *Variations* in Section 3). Footnotes are used throughout in referring to "more complex tasks", which are further extensions to the basic task not currently implemented in the TACTICS design or software.

The task: In TACTICS, a cognitive human (Blue defense) vs. computer agent (Red offense) game is played over a series of trials, in an area of interest, using data depicted on a Geographic Information System (GIS) display (see Figure 1). Each trial involves one point in a Blue region within the area of interest. [Note: In some *Variations* (Section 3), a trial may involve more than one point in the Blue region]. Red and Blue each have two options for action on a trial. Red may attack the Blue point, or else not attack. Blue may divert from the Blue point to avoid a possible Red attack, or else not divert and risk the consequence of a possible Red attack. The result of a showdown (Red attack and Blue not divert) is U units of utility won by Blue (lost by Red) at a probability P, or U units of utility won by Red (lost by Blue) at a probability 1-P. Blue loses 1 unit of utility when he diverts and Red does not attack, i.e., when Blue spends resources to divert and Red does not spend resources to attack. The outcome is 0 units of utility for Red and Blue when neither spends resources (i.e., Red does not attack and Blue does not divert), or when both spend resources but there is no showdown (i.e., Red attacks and Blue diverts). To minimize losses (i.e., optimize defense), Blue must acquire and apply knowledge of relevant probabilities and utilities. The Blue (human) player must also adapt to the outcomes of trials and detect changes in Red (agent) tactics. The task manipulates Blue (human) response demands as discussed in *Variations* (Section 3) to measure cognitive performance in ***inferencing*** over hypotheses, ***decision-making*** based on inferences, and ***foraging*** for new evidence.

The map: A GIS display (see Figure 1) outlines the region of Blue defense in an area of interest. In some *Variations*, a Blue player can "mouse click" to see ***"batch plots"*** of attacks over previous trials. A batch plot (Figure 2) is the cumulative display of significant activities (SIGACTS), i.e., attacks and outcomes that occurred over a series of trials, and can be "played-back" in time to show the trial-by-trial accumulation of SIGACTS.

A trial: On each trial (which represents a day in the area of interest), Blue receives a sequence of intelligence reports about spatial-temporal features of events in an attack script – see Table 2. The spatial features affect Red's *vulnerability* to Blue defense and *opportunity* to inflict damage. The temporal events include Red's latent *capability* to attack Blue and Red's latest *activity* near Blue points. Blue must first use these spatial and temporal clues in ***inferencing***, to estimate and update the probability that Red will attack on the current trial. Blue must then use the results of inferencing for ***decision-making***, to choose a Blue action (i.e., divert or ~divert) at the Blue point on the current trial. In *Variations* (Section 3) of the basic task, Blue also must make ***foraging*** decisions about where to obtain further information (at one of several Blue points), and perform forensic inferencing to diagnose Red tactics and detect changes in Red tactics.

OSINT: To start a trial, the location of planned Blue activity[1] is assumed to be reported in open-source media (OSINT), hence known by Red as well as Blue (see Figure 1). This is the location at which Red may potentially attack Blue on the current trial. The GIS also displays the shortest straight-line distance (r) measuring proximity of the Blue point to the border of the Blue region. A large distance implies a relatively large *vulnerability* for Red (and relatively small *vulnerability*) for Blue, if an attack is attempted by Red. Thus r affects the *probability* P that Blue will defeat Red if Red chooses to attack. This probability increases as r increases, per the function $P = 1 - e^{-vr}$ (see Figure 3). As such, P is the cumulative distribution function for a constant failure rate model corresponding to the exponential (Poisson) distribution (see Roberts, et al., 1981), which assumes that the probability of "failure" (i.e., Blue failure to defeat Red if Red attacks) is constant[2] for each delta-r in the integration performed to compute P. The value of P at the location is displayed by the GIS, and assumed known by both Blue and Red. Note that in TACTICS the value of P is always ≤ 0.5, see Section 4 *Solutions*. As such Blue is playing "defense" against Red, and the Blue objective is to minimize expected losses in a game where Blue's expected utility is ≤ 0.

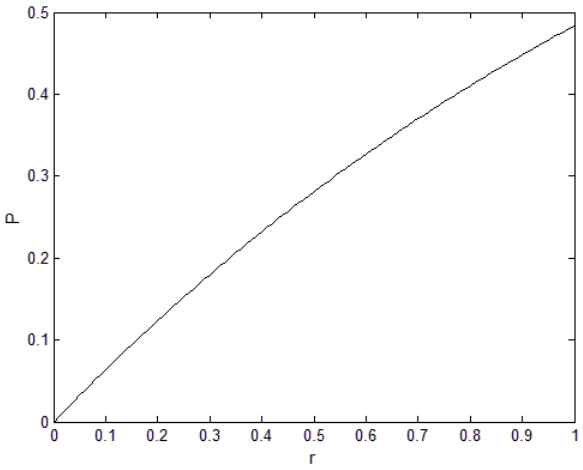

Figure 3: Probability that Blue will defeat Red if Red attacks a Blue point, as a function of distance (r) from the Blue point to the Blue border, $P = 1 - e^{-vr}$, assuming v = 2.

[1] The location(s) of Blue activity on each trial will be selected at random by the computer, much like a random "deal" of card(s) in poker. However, the random selection may be constrained by experimenters to ensure that stimuli are most suitable for assessment of human and model performance, see *Evaluation* (Section 5). A more complex task might allow the Blue player to choose Blue point(s) on each trial, subject to some appropriate constraints – such that both Blue and Red might anticipate the Blue points that are likely or not likely to be at stake on future trials.

[2] The failure rate v is assumed to be constant in time, although more complex tasks might vary v in space and time. A more complex task might also make P a function of further variables, besides just proximity (r), and/or might require that the Blue human (or a human teammate) estimate P as either a point estimate or a probability distribution. The Blue points and corresponding values of P are known by both Blue and Red, although more complex tasks could vary the availability and reliability of this knowledge between Red and Blue. Note that in the real world, estimating P from geospatial features of terrain might be considered a form of *suitability* analysis.

IMINT: Besides OSINT, both Red and Blue have access to open-source imagery intelligence (IMINT) showing buildings and other terrain features (see Figure 1). Of particular concern is the density of buildings in the vicinity of the Blue point, which is assumed to measure Red's *opportunity* to inflict damage on Blue. This opportunity is quantified as a utility (U), which is directly proportional to building density[3]. The value of U is computed and displayed by the GIS, and assumed known by both Blue and Red. Note that in TACTICS the values of U are limited to integers U = 2, 3, 4, or 5. This is to simplify the space of solutions for human experiments, see Section 4 *Solutions*.

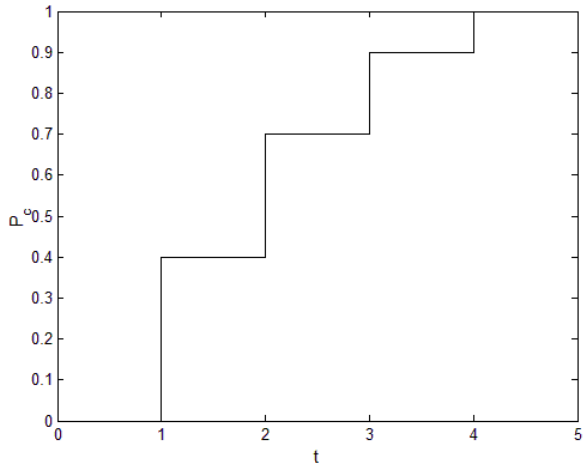

Figure 4: Probability that Red has the capability to attack, as a function of time (t = number of trials) since the last attack.

HUMINT: After OSINT (P) and IMINT (U), Blue receives an additional report from human intelligence (HUMINT), which is displayed by the GIS and seen only by Blue. This HUMINT report reflects Red's overall *capability* to recruit members, acquire weapons, transport forces, arrange escape, and satisfy other requirements for launching an attack against Blue. The capability to attack is modeled as a probability of attack, P_c, assuming Red wishes to attack (which depends on Red tactics, see BLUEBOOK below). More specifically, this Red capability (probability) P_c is 1.0 at the start of a mission and increases with time (t) after the last attack during the mission, per a discrete function that models temporal recency effects much like the continuous function P models spatial proximity effects. That is, P_c models temporal "failures" (i.e., Blue failure to prevent the Red attack *capability*) whereas P models spatial "failures" (i.e., Blue failure to defend, which is a Blue *vulnerability*)[4].

[3] A more complex task might make U a function of further spatial (and/or temporal) variables, besides just building density, and/or might require that the Blue human (or a human teammate) estimate U. Like P (from OSINT), the value of U on each trial (from IMINT) is known by both Red and Blue, although more complex tasks could vary the availability and reliability of this knowledge between Red and Blue. More complex tasks might also vary the subjective utility of objective utility U between Red and Blue, to simulate different value structures of asymmetric adversaries. Note that in the real world, estimating U from geospatial features of terrain might be considered a form of *suitability* analysis.

[4] The step function is assumed constant, although more complex tasks might vary the function with space and time. For example, in a more complex task Red's capability may depend on recent outcomes and their effects on the surrounding ("Green")

BLUEBOOK: Besides HUMINT, Blue is asked to consider the Blue "team" knowledge about Red's tactics, as expressed in a Blue handbook called the BLUEBOOK. In particular, the BLUEBOOK specifies how Red's *propensity* to attack, given the *capability* to attack, depends on *vulnerability* (i.e., probability P discussed under OSINT above) and *opportunity* (i.e., utility U discussed under IMINT above). In *Variations* (Section 3) of the basic task, the BLUEBOOK may represent Red's tactics for one or more Red "styles" and the style may be known or unknown[5]. After referring to the BLUEBOOK, **Blue is asked to report his belief about Red's propensity to attack**, i.e., the likelihood (measured by a conditional probability) P_p that Red will attack on the current trial, assuming that Red has the capability to attack on the current trial. Then, after reporting P_p, **Blue is asked to adjust his estimate of the probability that Red will attack on the current trial**, considering the HUMINT value of P_c as well as Blue's own report of P_p. The answer, $P_{p,c}$, represents Blue's best estimate of Red's attack probability based on intelligence about Red's *propensity* to attack and Red's *capability* to attack.

SIGINT: Finally, after reporting $P_{p,c}$, Blue receives a report from signals intelligence (SIGINT) about Red *activity* on the current trial. This report is based on communications (e.g., cell phone usage) that would signal Red coordination around the Blue point to support a Red attack. The SIGINT report is always of limited reliability, because SIGINT will sometimes "miss" Red attack signals and sometimes (but less likely) "hear" Red attack signals when none exist. Also, it is assumed that SIGINT detects only the occurrences of communications and not the contents of those communications. More specifically, if ground truth is "yes" (i.e., Red is actually coordinating an attack) then "YES" will be reported by SIGINT at 60% probability and "NO" will be reported by SIGINT at 40% probability. On the other hand, if ground truth is "no" (i.e., Red is not actually coordinating an attack) then "NO" will be reported by SIGINT at 80% probability and "YES" will be reported by SIGINT at 20% probability. In experimental manipulations (see *Variations*, Section 3), SIGINT resources may be limited such that Blue must choose a location at which to collect SIGINT. The SIGINT return and associated likelihoods[6] can be used to infer the probability P_t of Red's *activity* on the current trial, independent of Red's *propensity* to attack (see P_p above) and independent of Red's *capability* to attack (see P_c above).

---

population – e.g., P_c may increase when Red is winning and decrease when Red is losing. Also in more complex tasks, the estimating of Red's capability P_c (either point estimate or a probability distribution) might be performed by a Blue human (or a human teammate). Finally, in more complex tasks, the planning, transport, and other precursors (or successors) to Red attack (and Blue defense) may be treated explicitly as separate events, and modeled with conditional probabilities that relate these events to each other (and spatial context) in Red tactics. Note that in the real world, these *capability* analyses might be performed in conjunction with *suitability* analyses, like those corresponding to P (vulnerability) and U (opportunity) mentioned above.

[5] For example, the BLUEBOOK might specify the propensity function $P_p = fn(P, U)$ by which Red makes his choice to attack (or ~attack), given Red capability to attack, for a "Passive" style and for an "Aggressive" style. In that case Blue would need to infer the likelihood of each Red style in order to estimate $P_p = fn(P, U)$ from the BLUEBOOK values.

[6] More complex tasks might vary the nature of SIGINT reports, i.e., to include the contents of messages as well as their probabilities, and/or to reflect a variable area around the Blue point, and/or to vary the reliability of SIGINT with spatial-temporal context.

After receiving SIGINT, **Blue is asked[7] to report the probability of attack based only on SIGINT (P_t). Blue is then asked to update his estimate of the probability that Red will attack on the current trial ($P_{t,p,c}$)**, i.e., considering the likelihood of Red's activity (per P_t) as well as the prior combination of Red's propensity and capability ($P_{p,c}$). This yields a final estimate of the Red attack probability[8] $P_a = P_{t,p,c}$.

Red's move: Red's action, chosen without Blue knowing, is either to "attack" or "~attack" (not attack). Assuming Red chooses to attack, the attack will actually occur only if Blue does not foil the attack by a move to "divert" Blue forces away from the Blue point at stake. Red's choice on the trial (R_t) depends on Red's *capability* P_c (given by HUMINT) and *propensity* P_p (given by BLUEBOOK), where P_p depends on *vulnerability* (P given by OSINT) and *opportunity* (U given by IMINT). The propensity P_p (per BLUEBOOK, see above) is reflected in Red's tactics (R_R), which in turn reflect the reward structure (discussed below) by which outcomes are scored. In general, Red's propensity to attack would also depend on Red's beliefs about Blue's tactics (R_B), because the expected utility of Red's action depends on the probability $P_{\sim d}(R_B)$ that Blue will not divert forces and hence will face a potential attack. However the current TACTICS assumes that Red's tactics are not dependent on Blue's tactics, i.e., Red's tactics are only a function of P, U, and time since the last attack.

Blue's move: Blue's action, chosen without Red knowing, is either to "divert" or "~divert" (not divert). Blue's choice on the trial (B_t) is governed by his *intentionality (rationality)* and reflects Blue tactics (B_B), which depend on *vulnerability* (P) and *opportunity* (U) as well as Blue's beliefs about the probability $P_a(B_R)$ of Red attack. Note that this probability in turn depends on Blue's model of Red tactics, B_R. The Red tactics are known for some missions, but for other missions the Red tactics are unknown (hence B_R must be inferred by Blue). After reporting his estimate of Red's attack probability $P_a(B_R) = P_{t,p,c}$ (see SIGINT above), **Blue is asked to choose an action, either "divert" or "~divert"**. This Blue choice is based on knowledge of the "payoff matrix" (see Figure 5), which is also known by Red, and which specifies the expected utility to be gained or lost by each player (Blue and Red) for each possible combination of Blue-Red actions: (~d, a), (~d, ~a), (d, a), and (d, ~a).

[7] Note that this and other questions may not be asked on every trial of every mission. For example, the answer to the question here (P_t) would be the same or similar across trials for each value of SIGINT ("YES" or "NO"), as long as the SIGINT reliabilities are held constant.

[8] In the real world, estimating the probability P_a of Red attack (along with estimating the probability P of Blue success and utility U of the target) is analogous to *TTP (Tactics, Techniques, and Procedures)* analysis. This type of analysis integrates various suitability and activity analyses, along with historical and inferential knowledge about enemy tactics, to produce *actionable* I&W (Indications and Warnings, see Grabo, 2004) intelligence estimates such EMPCOA (Enemy's Most Probable Course of Action) and EMDCOA (Enemy's Most Dangerous Course of Action). In the real world, these intelligence estimates are relayed to and employed by operational forces. In TACTICS, Blue is playing the role of both intelligence and operations, as he uses his own *inferences* (e.g., P_a) to make his own *decisions* (see "Blue's move"). More complex experiments could involve a team of two (or more) Blue players, i.e., separating the intelligence and operations functions in order to investigate communication and coordination in team sensemaking. Likewise more complex experiments could involve a team of Blue analysts, each performing one or more of the various suitability (P, U), capability (P_c,), propensity (P_p), activity (P_t), or intentionality (P_a) analyses.

	a (Red attack)	~a (Red ~attack)
~d (Blue ~divert)	U * [2 * P - 1]	0
d (Blue divert)	0	-1

	a (Red attack)	~a (Red ~attack)
~d (Blue ~divert)	-U * [2 * P - 1]	0
d (Blue divert)	0	+1

Figure 5: Payoff matrix for Blue (top matrix) and Red (bottom matrix).

The score: After Blue's move, Red's move is revealed and the values of P and U are used to generate a significant activity (SIGACT) report of the outcome. Referring to Figure 5, in the case of a showdown (i.e., Red attack and Blue ~divert) one of two outcomes [+U Blue (-U Red), -U Blue (+U Red)] is randomly chosen by the computer at probabilities [P, 1-P], respectively. This produces expected utilities as indicted in the upper-left cell (~d, a) of each payoff matrix (Blue and Red) above. For all other combinations of actions, i.e., (~d, ~a), (d, a), and (d, ~a), the payoff is a fixed value. Note that the payoffs for Blue and Red are always equal in magnitude but opposite in sign, so TACTICS is a zero-sum game[9].

A batch: A "batch" is a series of trials, with each trial involving a new Blue point in the region of Blue defense – i.e., on the same GIS map (Figure 1). The parameters of Red's tactics are held constant over trials of Missions 1-3 (see *Variations*, Section 3). In Missions 4-5, Red tactics will change at some point in the mission, and the Graphical User Interface (GUI) allows Blue to make "batch plots" (see Figure 2) in order to diagnose the Red tactics and detect the changes in Red tactics.

[9] This scoring system rewards a player (Blue or Red) with utility +U for winning a showdown, which occurs when Red attacks and Blue ~divert. The utility is 0 for Blue (0 for Red) if Red attacks and Blue diverts; also 0 for Blue (0 for Red) if Red ~attack and Blue ~divert. The utility is -1 Blue (+1 Red) when Red ~attack and Blue diverts, because Blue invested resources in the divert and Red did not invest resources in an attack. More complex tasks might use other scoring systems, including non-zero-sum utilities for Red and Blue to reflect the relative importance of various outcomes to asymmetric adversaries. More complex tasks might also make other aspects of the game state dependent on outcomes, e.g., changing the Blue border in response to Blue wins (growing the Blue region) or Red wins (shrinking the Blue region), and/or changing various other parameters (e.g., v in the *vulnerability* model) in response to Blue or Red wins.

3 Variations

The basic task (see *Description*, Section 2) is manipulated across *missions* as needed to measure Blue sensemaking processes and cognitive biases (per BAA Table 3). In particular, it is useful to distinguish three different but related cognitive processes as follows: ***inferencing***, ***decision-making*** (based on inferencing), and ***foraging*** **(**based on inferencing and decision-making). These processes are highlighted and evaluated in Missions 1-3 as described below. In addition is it useful to distinguish between *prognostic inferencing*, to predict future attacks, and *forensic inferencing*, to explain previous attacks. Missions 1-3 are focused on prognostic inferencing, whereas Missions 4-5 require forensic inferencing as a basis for prognostic inferencing.

Mission 1. You judge the chance *(inferencing)*: Mission 1 is focused on measuring how Blue updates his HUMINT and BLUEBOOK prior ($P_{p,c}$) with SIGINT (P_t) likelihoods to compute a posterior probability $P_{t,p,c}$. Mission 1 also measures how Blue combines P_p from BLUEBOOK with P_c from HUMINT to compute the prior $P_{p,c}$. Each trial of Mission 1 involves only one Blue point, and the Red tactics (P_p) are specified by the BLUEBOOK as a function of P and U, P_p = fn(P, U), as follows:

	U = 2 or 3	**U = 4 or 5**
P > 25%	20%	40%
P ≤ 25%	60%	80%

Based on previous research (Burns, 2007) and pilot studies, we expect to see a conservative bias in human posteriors $P_{t,p,c}$, where $P_{t,p,c}$ is computed as an average of P_t and $P_{p,c}$ rather than a Bayesian-normalized product of P_t and $P_{p,c}$. This bias can be characterized ***Anchoring and Adjustment*** (Tversky & Kahneman, 1974), where P_t and $P_{p,c}$ act as anchors and the averaging of these anchors reflects an inadequate adjustment made in computing the posterior $P_{t,p,c}$. We also expect to see a conservative bias in estimates of P_t itself. This bias can be characterized as ***Availability*** (Tversky & Kahneman, 1974), where humans tend to use the readily available SIGINT likelihood P(SIGINT| attack) as a surrogate for the Bayesian-normalized posterior P_t = P(attack| SIGINT). Finally, we expect to see a bias in human estimates of the prior $P_{p,c}$. This bias can be characterized as a form of ***Representativeness*** known as the "conjunction fallacy", whereby humans compute $P_{p,c}$ as an average of P_p and P_c, and thereby fail to compute a joint probability $P_{p,c} = P_p * P_c$ that is less than P_p and less than P_c.

As such, Mission 1 addresses Octaloop (Burns, 2014) step [3] estimating likelihoods as well as Octaloop step [4] aggregating confidence. Note that here in Mission 1, Blue's choice to "divert" or "~divert" will be made by a Blue agent (not the human), to ensure that all human subjects receive the same post-judgment stimuli (which may affect Blue's inferencing behavior).

Mission 1 addresses the BAA "core sensemaking processes" of ***Learn Frames (Features)***, ***Recognize Patterns / Select a Frame, Assess the Frame, Re-frame (Features).***

Mission 2. You make the choice *(decision-making)*: Mission 2 is focused on measuring how Blue uses his estimate of $P_a = P_{t,p,c}$ from inferencing (discussed above), along with the known values of P and U, to make choices (Octaloop step [5], speculating consequences) of "divert" or "~divert" and then adapt to outcomes (Octaloop step [6], evaluating consequence). Like Mission 1, Mission 2 also measures inferences of $P_{p,c}$ and $P_{t,p,c}$. Each trial involves only one Blue point, but the Red tactics are not known for certain. Instead, the BLUEBOOK specifies attack probabilities P_p as a function of P and U, for two Red styles: Passive and Aggressive.

The Passive Red tactics, P_p(Passive) = fn(P, U), are as follows:

	U = 2 or 3	**U = 4 or 5**
P > 25%	20%	30%
P ≤ 25%	40%	50%

The Aggressive Red tactics, P_p(Aggressive) = fn(P, U), are as follows:

	U = 2 or 3	**U = 4 or 5**
P > 25%	50%	60%
P ≤ 25%	70%	80%

Using these two BLUEBOOK tables, a normative solution for P_p can be can be computed on each trial using the attack history up to that trial, see *Forensic Inferencing* in Section 4.4. A normative solution for each stage of *Prognostic Inferencing*, per Section 4.1, can then be computed in the same manner as for Mission 1. Finally, given the results of inferencing, a normative solution for *Decision-making* (see Section 4.2) computes the Blue option (divert or ~divert) with highest expected utility. We expect that humans will exhibit a form of ***Probability Matching*** (Burns & Demaree, 2009) in which choices to divert or ~divert are biased, such that human decisions will often deviate from normative decisions.

In addition to the core processes addressed in Mission 1, Mission 2 addresses the BAA "core sensemaking processes" of ***Learn Frames (Actions)***, ***Generate Expectations of Missing Data (SIGACT), Acquire Additional Data (SIGACT), Re-frame (Actions)***.

Mission 3. You send the spies *(foraging)*: Mission 3 is focused on measuring how Blue allocates limited resources in collecting information (per Octaloop step [7], anticipating evidence) to support choices of actions (divert or ~divert) like those made in Mission 2 (per steps [5] and [6] of Octaloop). Each trial involves two Blue points, but Red can attack at only one (or neither) of the Blue points. Also, Blue can obtain a SIGINT[10] report at only one of the two points.

A normative solution for Blue's choice of SIGINT location (Section 4.3.2) can be computed by considering both SIGINT options (point 1 and point 2), and by evaluating the expected gain in information from each option. Before SIGINT, Blue is asked to consider Red's *propensity* P_p (given by the BLUEBOOK) and Red's *capability* P_c (given by HUMINT) in order to estimate $P_{p,c}$ without SIGINT. After reporting $P_{p,c}$, Blue is asked to pick one Blue point for collecting SIGINT, before making his decision to divert or ~divert at each point.

For example, Blue may choose to get SIGINT at the Blue point of highest Red attack probability (highest $P_{p,c}$), or the point with highest Blue vulnerability (lowest P), or the point of highest utility (highest U). We expect to see ***Confirmation Bias in Seeking Evidence*** (Nickerson, 1998; Klayman & Ha, 1987; Fischhoff & Beyth-Marom, 1983), where Blue seeks SIGINT on the Blue point with the highest attack probability. However, as noted in Section 4.3.2, this so-called bias is actually the optimal behavior for maximizing expected information gains from SIGINT. Therefore, the non-normative bias is to NOT always seek SIGINT at the location with highest $P_{p,c}$, and the frequency at which humans exhibit this behavior will be taken as a measure of ***Confirmation Bias***.

In addition to the core processes addressed in Missions 1 and 2, Mission 3 addresses the BAA "core sensemaking processes" of ***Generate Expectations of Missing Data (SIGINT), Acquire Additional Data (SIGINT)***.

Missions 4,5. You spot the change: Missions 4-5 differ from Missions 1-3 in that Red tactics change at some point in time. In Mission 4, the change is from Passive to Aggressive, or vice versa, where the parameters of each style are the same as in Mission 2 above. For Mission 5, one style is P-sensitive, as defined by the following values of P_p(P-sensitive):

	U = 2 or 3	**U = 4 or 5**
P > 25%	40%	40%
P ≤ 25%	60%	60%

The other style is U-sensitive, as defined by the following values of P_p(U-sensitive):

[10] More complex tasks might present more than two Blue points on each trial, and/or or require that the Blue player choose among various INTS (i.e., OSINT, IMINT, HUMINT, SIGINT) with the choice being subject to some specified constraint(s) – e.g., choose only one or two or three of the four INTS, and do so at only some (not all) of the Blue points.

	U = 2 or 3	**U = 4 or 5**
P > 25%	20%	80%
P ≤ 25%	20%	80%

In these missions, Blue must infer Red's tactics in the first place, as well as detect the change at some unknown point in time, in order to support inferencing and decision-making. To enable testing of more trials, the sources of intelligence for Missions 4-5 are limited to OSINT and IMINT (i.e., no HUMINT or SIGINT) "within" each trial. Also, to support Blue's inferences about Red tactics "between" trials, on selected trials (e.g., every ten trials) Blue is allowed to create and inspect "batch plots" of past attacks. In so doing a player is performing ***forensic foraging*** through previous attack histories (SIGACTS), which differs from the *prognostic foraging* for intelligence (SIGINT) in Missions 1-3.

Missions 4 and 5 differ from one another primarily in the difficulty of detecting Red tactics and the change in Red tactics. In Mission 4, Red's tactics are known to reflect either a "Passive" or "Aggressive" style, and the style can be inferred from the total frequency of past attacks. In Mission 5, the possible Red styles are "P-sensitive" or "U-sensitive", and these styles cannot be inferred only from the total frequency of past attacks. Instead, the inference requires attention to values of P and U in subsets of past attacks.

Missions 4-5 are designed to measure three final biases, namely ***Change Blindness***, ***Persistence of Discredited Evidence***, and ***Satisfaction of Search***. For ***Change Blindness***, we expect that humans will be delayed in detecting the change of Red tactics, and possibly even fail to detect the change at all – especially in Mission 5. For ***Persistence of Discredited Evidence***, we expect that human uncertainty about the Red style will persist to the end of Mission 4, i.e., even after obtaining ample evidence (SIGACTS) to discredit beliefs held before the change in Red style. For ***Satisfaction of Search***, we expect that humans will terminate their searches for data through batch plots prematurely, i.e., not perform an exhaustive search through all past attacks that are available in batch plots.

Missions 4-5 address Octaloop steps [8] discriminating evidence, [1] isolating evidence, and [2] generating hypotheses.

In addition to the core processes addressed in Missions 1-3, Missions 4-5 address the BAA "core sensemaking processes" of ***Learn Frames (Tactics), Generate Expectations of Missing Data (Batch Plots), Acquire Additional Data (Batch Plots), Re-frame (Tactics).***

4 Solutions

4.1 Inferencing (Prognostic)

4.1.1 Inferencing at One Blue Point

For *inferencing* in a *prognostic* sense, i.e., to predict the probability of Red attack, the normative solution at each stage of a trial depends on the probabilities being aggregated. For Mission 1, P_p is given by the BLUEBOOK based on OSINT (P) and IMINT (U), and P_c is given by HUMINT. For Mission 2, Blue must perform *forensic inferencing* (see Section 4.4) to obtain the value of P_p. In both missions, P_c and P_p are normatively combined as a simple product because $P_{p,c}$ = P(propensity, capability) = P(capability) * P(propensity|capability) = P_c * P_p.

In the next stage of a trial, $P_{p,c}$ and P_t are normatively combined in a Bayesian update: $P_{t,p,c} \sim P_t * P_{p,c}$ and $(1\text{-}P_{t,p,c}) \sim (1\text{-}P_t) * (1\text{-}P_{p,c})$, where ~ implies a normalization (i.e., division by the sum $[P_t * P_{p,c} + (1\text{-}P_t) * (1\text{-}P_{p,c})]$ to ensure that the posteriors $P_{t,p,c}$ and $1\text{-}P_{t,p,c}$ sum to 1). Notice that aggregation at this stage is different than at the first stage, because here at the second stage the probabilities being combined are both referring to the same hypothesis that may or may not be true, namely the hypothesis that Red will attack. Conversely, at the first stage, the probabilities being combined refer to different hypotheses, namely a hypothesized *capability* to attack (P_c) and a hypothesized *propensity* to attack (P_p) assuming the capability, where an actual attack would require that both hypotheses be true.

In Missions 1 and 2, another twist arises because P_t is not provided directly but rather must be inferred from the SIGINT likelihoods (Burns, 2006). These likelihoods are given to Blue as follows: P(Y|y) = 60%, P(N|y) = 40%, P(Y|n) = 20%, and P(N|n) = 80%, where "Y" and "N" refer to signals (SIG = YES or NO) whereas "y" and "n" refer to the ground truth (yes or no). In effect, the human must first "invert" the SIGINT likelihoods from P(evidence|hypothesis) to compute posteriors P(hypothesis|evidence) using Bayes Rule. This yields P_t = P(y|S) and $1\text{-}P_t$ = 1-P(y|S) = P(n|S) for whichever signal was received (S = Y or S = N). For example, if SIGINT reports Y then we have (assuming a uniform prior):

$$P_t = P(y|Y) = P(Y|y) / [P(Y|y) + P(Y|n)] = 60\% / [60\% + 20\%] = 75\%$$

$$1\text{-}P_t = P(n|Y) = 25\%.$$

On the other hand, if SIGINT reports N then we have (assuming a uniform prior):

$$P_t = P(y|N) = P(N|y) / [P(N|y) + P(N|n)] = 40\% / [40\% + 80\%] = 33\%$$

$$1\text{-}P_t = P(n|N) = 67\%.$$

In short, the Bayesian value of P_t is 75% (not 60%) if SIGINT reports Y, and 33% (not 40%) if SIGINT reports N.

4.1.2 Inferencing at Two Blue Points

The above solutions for prognostic inferencing apply to trials of Missions 1 and 2, where all INTS (OSINT, IMINT, HUMINT, SIGINT) are provided at only one Blue point (i.e., one location in the region of Blue defense) on each trial. In Mission 3, each trial presents INTS at two Blue points. The same solution for $P_{p,c} = P_c * P_p$ applies at each location on a trial of Mission 3, because the HUMINT (P_c) representing Red attack *capability* applies equally to any and all locations. However, in Mission 3 the BLUEBOOK specifies different values for Red attack *propensity* (P_p) at each location based on OSINT (P) and IMINT (U), as follows:

	U = 2 or 3	**U = 4 or 5**
P > 25%	10%	20%
P ≤ 25%	30%	40%

Note that each of these values is one half the corresponding value specified by the BLUEBOOK in Mission 1, because here in Mission 3 Red may attack at either (or neither) of the two Blue locations.

After reporting $P_{p,c}$ on a trial of Mission 3, at each of two Blue locations, Blue must choose a location (denoted 1 or 2) at which to receive SIGINT. The normative solution for this decision is developed in Section 4.3, *Foraging*. Depending on whether SIGINT returns "chatter" (SIG = YES) or "silence" (SIG = NO), the Bayesian distribution $\{P_t, 1\text{-}P_t\}$ at the location where SIGINT was obtained (call it location 1) will be either {75%, 25%} or {33%, 67%}, see Section 4.1.1.

Because Red can attack at only one (or neither) location, but not both locations, there are three hypotheses {A, B, C} that must be considered: A = attack at location 1; B = attack at location 2; C = no attack at location 1 or 2. The priors are given by $\{P_{p,c,1}, P_{p,c,2}, 1\text{-}P_{p,c,1}\text{-}P_{p,c,2}\}$, respectively. The likelihoods given "chatter" at location 1 (assumed to be the location at which SIGINT was obtained) are {75%, 12.5%, 12.5%}, and the likelihoods given "silence" at location 1 are {33%, 33.5%, 33.5%}. Note that these likelihood distributions are each of the form $\{P_{t,1}, (1\text{-}P_{t,1})/2, (1\text{-}P_{t,1})/2\}$, because the probability $1\text{-}P_{t,1}$ applies to hypotheses B and C (i.e., ~A).

Finally, the prior distribution is updated using the likelihood distribution, to compute the posterior distribution as a Bayesian-normalized product of prior and likelihood. Note that the posterior probability of attack will differ from the prior probability of attack even at location 2 for which no SIGINT was obtained. This is because of the dependency between locations introduced by the assumption that Red can attack at only one (or neither) location.

4.2 Decision-Making

4.2.1 Non-equilibrium Solution

The non-equilibrium solution for each player (Blue or Red) is computed from the payoff matrix (Figure 5) by assuming that the probability of an opponent's action is known.

For Blue, the expected utility (E) of each option (divert or ~divert) is computed as follows:

$$E_d = P_a * \{0\} + (1 - P_a) * \{-1\} = P_a - 1$$
$$E_{\sim d} = P_a * \{U * [2 * P - 1]\} + (1 - P_a) * \{0\} = P_a * \{U * [2 * P - 1]\}$$

where P_a is the probability that Red will attack, P is the probability that Blue will defeat Red if Red attacks, and U is the utility gained by the winner of a showdown.

Blue should divert if $E_d > E_{\sim d}$, i.e., if $E_d - E_{\sim d} > 0$:

$$E_d - E_{\sim d} = (P_a - 1) - P_a * \{U * [2 * P - 1]\} > 0$$
$$= P_a - (P_a * U * 2 * P) + (P_a * U) > 1$$

or $-P * [2 * U * P_a] > -P_a * (U + 1) + 1$.

<u>Hence Blue should divert when:</u> $P < [P_a * (U + 1) - 1] / (2 * U * P_a)$.

For Red, the expected utility (E) of each option (attack or ~attack) is computed as follows:

$$E_a = -P_{\sim d} * \{U * [2 * P - 1]\} + (1 - P_{\sim d}) * \{0\} = -P_{\sim d} * \{U * [2 * P - 1]\}$$
$$E_{\sim a} = P_{\sim d} * \{0\} + (1 - P_{\sim d}) * \{1\} = 1 - P_{\sim d}$$

where $P_{\sim d}$ is the probability that Blue will ~divert, P is the probability that Blue will defeat Red if Red attacks, and U is the utility gained by the winner of a showdown.

Red should attack if $E_a > E_{\sim a}$, i.e., if $E_a - E_{\sim a} > 0$. That is:

$$E_a - E_{\sim a} = -P_{\sim d} * \{U * [2 * P - 1]\} - (1 - P_{\sim d}) > 0.$$

Notice this is the same as the equation for Blue, if we replace P_a (in the equation for Blue) by $P_{\sim d}$ (in the equation for Red).

<u>Thus Red should attack when:</u> $P < [P_{\sim d} * (U + 1) - 1] / (2 * U * P_{\sim d})$.

These non-equilibrium solutions for Blue and Red are illustrated in Figure 6 and discussed further below for two cases of interest: $P > 0.5$ and $P < 0.5$.

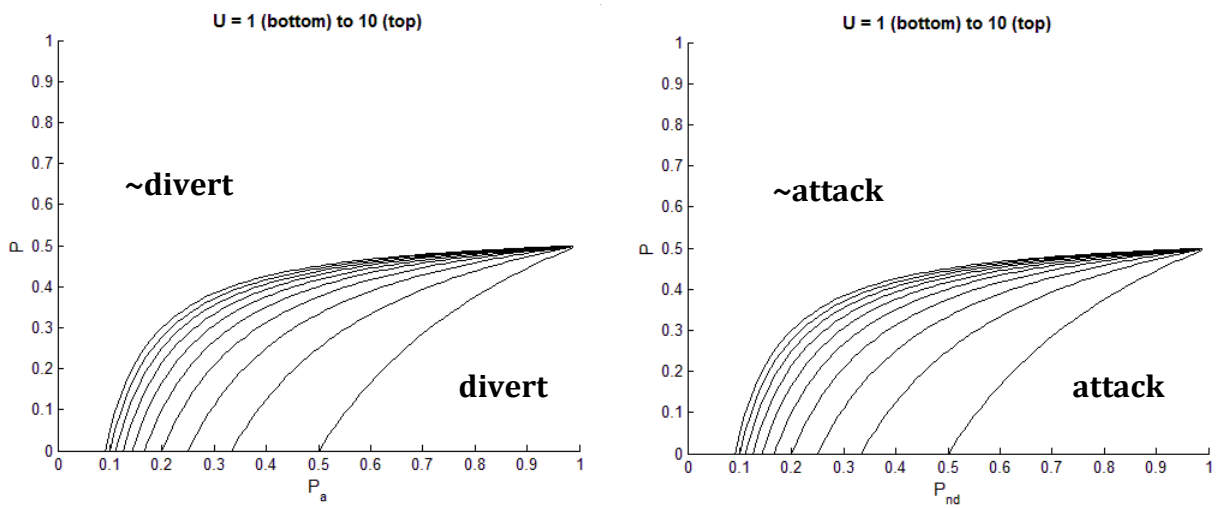

Figure 6: Non-equilibrium solutions for Blue (left) and Red (right).

For Blue, when P > 0.5: If Blue ~divert then his expected utility is > 0 (if Red attacks) or = 0 (if Red ~attack). If Blue diverts then his expected utility is = 0 (if Red attacks) or = -1 (if Red ~attack). Thus regardless of P_a, Blue should always ~divert when P > 0.5.

For Blue, when P < 0.5: If Blue ~divert then his expected utility is < 0 (if Red attacks) or = 0 (if Red ~attack). If Blue diverts then his expected utility is = 0 (if Red attacks) or -1 (if Red ~attack). Because neither option (~divert or divert) is always better, Blue must consider the probability P_a of Red attack. As P_a decreases, ~divert by Blue is less likely to result in a showdown with negative expected utility and more likely to result in 0 expected utility. Thus, the P threshold for ~divert decreases (from 0.5 to smaller values) as P_a decreases (from 1 to smaller values) along a line of constant U (see Figure 6). At a given value of P_a, the expected loss (i.e., magnitude of expected utility < 0) resulting from Blue ~divert and Red attack increases as U increases. Thus the P threshold for ~divert increases as U increases.

For Red, when P > 0.5: If Red attacks then his expected utility is < 0 (if Blue ~divert) or = 0 (if Blue diverts). If Red ~attack then his expected utility is = 0 (if Blue ~divert) or = +1 (if Blue diverts). Thus regardless of $P_{\sim d}$, Red should always ~attack when P > 0.5.

For Red, when P < 0.5: If Red attacks then his expected utility is > 0 (if Blue ~divert) or = 0 (if Blue diverts). If Red ~attack then his expected utility is = 0 (if Blue ~divert) or +1 (if Blue diverts). Because neither option (attack or ~attack) is always better, Red must consider the probability $P_{\sim d}$ of Blue ~divert. As $P_{\sim d}$ decreases, attack by Red is less likely to result in a showdown with positive expected utility and more likely to result in 0 expected utility. Thus, the P threshold for ~attack decreases (from 0.5 to smaller values) as $P_{\sim d}$ decreases (from 1 to smaller values) along a line of constant U (see Figure 6). At a given value of $P_{\sim d}$, the expected gain (i.e., magnitude of expected utility > 0) resulting from Red attack and Blue ~divert increases as U increases. Thus the P threshold for ~attack increases as U increases.

4.2.2 Nash-Equilibrium Solution

The Nash-equilibrium solution is computed, in two steps (Davis, 1997), from the payoff matrix in Figure 5. Note that this solution applies only to a zero-sum game.

First, for the case where P > 0.5, inspection of the payoff matrix (Figure 5) shows that ~divert (~d) dominates divert for Blue and ~attack (~a) dominates attack (a) for Red. Thus when P > 0.5 Blue should always ~divert (~d) and Red should always ~attack (~a). Also see Figure 6 above. The "value" of the game to each player is the expected utility assuming Blue always chooses ~divert and Red always chooses ~attack. This value, per the payoff matrix, is 0 for Blue and Red.

Then, for the case where P < 0.5, the optimal strategy for each player is a ***"mixed strategy"*** where each option is played at a probability ($P_{\sim d}$ for Blue and P_a for Red), which in turn depends on P and U.

For Blue, we consider the expected utility (across options, divert and ~divert) for each of Red's options (i.e., attack or ~attack). If Red attacks, Blue's expected utility is:

$$P_{\sim d} * \{U * [2 * P - 1]\} + (1 - P_{\sim d}) * \{0\}$$

where $P_{\sim d}$ is the probability that Blue will ~divert and $1 - P_{\sim d}$ is the probability that Blue will divert. If Red ~attack, Blue's expected utility is:

$$P_{\sim d} * \{0\} + (1 - P_{\sim d}) * \{-1\}$$

Because the game is zero-sum, Red's expected utility for each Red action is always the negative of Blue's expected utility (derived above). Therefore Blue's mixed strategy ($P_{\sim d}$) can be computed by equating the two expected utilities written above and solving for $P_{\sim d}$ as follows:

$P_{\sim d} * \{U * [2 * P - 1] = P_{\sim d} - 1$, which reduces to:

$P_{\sim d} * [2 * P * U - U - 1] = -1$

So Blue's optimal mixed strategy is as follows (see Figure 7): $\mathbf{P_{\sim d} = 1 / [1 - U * (2 * P - 1)]}$.

Using the same approach to solve for Red's optimal mixed strategy we obtain:

$$-P_a * \{U * [2 * P - 1]\} + (1 - P_a) * \{0\} = P_a * \{0\} + (1 - P_a) * \{1\}$$

This produces an equation for P_a that is the same as the equation for $P_{\sim d}$ above.

So Red's optimal mixed strategy is as follows (see Figure 7): $\mathbf{P_a = 1 / [1 - U * (2 * P - 1)]}$.

Finally, the value of the game to a player is the expected utility for either option (e.g., Blue ~divert or divert) assuming the numerical value of the associated mixed strategy. Thus the value of the game for Blue is given by $P_{\sim d} - 1$, as follows:

$V_B = P_{\sim d} - 1 = 1 / [1 + U - 2 * U * P] - 1$
$= [1 - 1 - U + 2 * U * P] / [1 + U - 2 * U * P].$

<u>The value of the game for Blue is (see Figure 8):</u> V_B = (U * [2 * P - 1]) / (1 - U * [2 * P - 1]).

<u>The value of the game for Red is (see Figure 8):</u> V_R = -(U * [2 * P - 1]) / (1 - U * [2 * P - 1]).

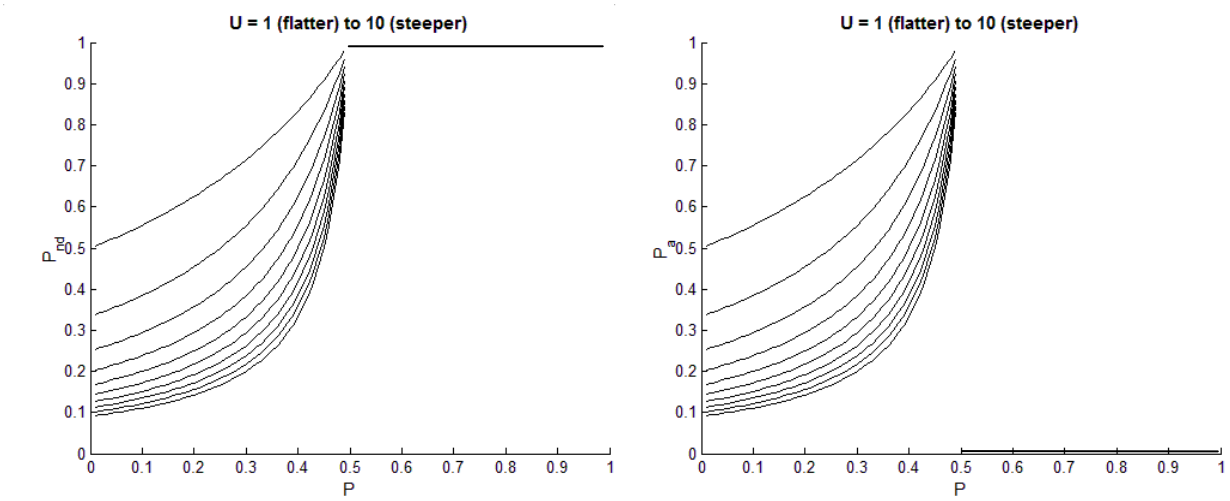

Figure 7: Nash-equilibrium solutions for Blue (left) and Red (right).

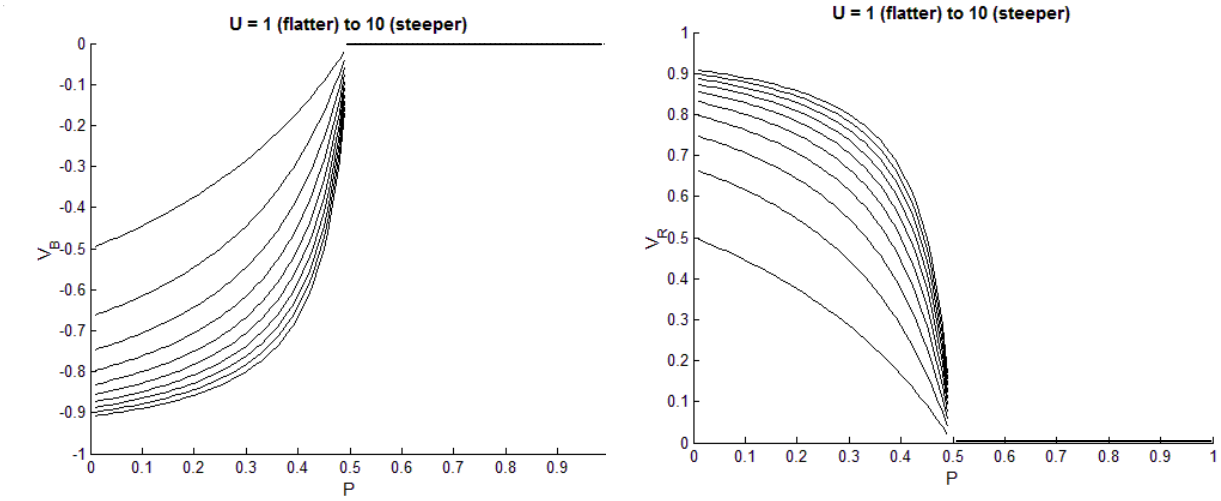

Figure 8: Nash-equilibrium value of the game to Blue (left) and Red (right).

4.2.3 Decision-Making at Two Blue Points

When two (or more) Blue points appear on a trial, Blue must consider the possibility of Red attack at each point. As discussed in Mission 3 of *Variations* (Section 3), Red can attack at only one (or neither) Blue point, but Blue can divert (or ~divert) from either one or both Blue points.

Thus Blue has four options across the points [1,2] as follows: A = [d_1,d_2], B = [d_1,~d_2], C = [~d_1,d_2], and D = [~d_1,~d_2]. And Red has three options across the points [1,2] as follows: A = [a_1,~a_2], B = [~a_1,a_2], and C = [~a_1,~a_2].

Unlike the simpler case where Blue only needs to consider the Red attack probability P_a at one point, Blue now has to estimate the probability of each Red option (A, B, C). The non-equilibrium strategy is described here, assuming the probability of each Red option (which depends on Blue's model B_R of Red tactics) is specified in the BLUEBOOK. To start, it is assumed that the BLUEBOOK specifies a "two point" propensity function $P_p = fn(P, U, B_R)$ that can be computed for each of the two Blue points, P_{p1} and P_{p2}, using the probabilities (P_1, P_2) and utilities (U_1, U_2) at these two points as known from OSINT and IMINT, respectively. Note that $P_{p1} + P_{p2} \leq 1$, because Red can attack at only one (or neither) Blue point.

Each value of P_p can then be combined with P_c (which is the same for each Blue point) and P_t (see *Foraging* below), to compute the probability of attack at each point: $P_{a1} = P_{t,p,c,1}$ and $P_{a2} = P_{t,p,c,2}$; also the probability of no attack, which is equal to $1-P_{a1}-P_{a2}$. This gives Blue the probability of each Red option (A, B, C).

Using these three probabilities, Blue can use the payoff matrix along with known values of probabilities (P_1, P_2) and utilities (U_1, U_2) to compute the expected utility for each Blue option: A = [d_1,d_2], B = [d_1,~d_2], C = [~d_1,d_2], and D = [~d_1,~d_2]. Given the resulting vector of expected utilities [U_A, U_B, U_C, U_D], the optimal Blue decision is to always choose the option with the highest expected utility. Unlike the simpler case of one Blue point analyzed in Section 4.2.1, the optimal solution in this case is a more complex function of three (not just two) Red probabilities and four (not just two) Blue options – hence not readily illustrated in parametric plots like Figure 6.

4.3 Foraging

For *foraging*, in Mission 3, Blue must choose one of two Blue points at which to receive SIGINT. After SIGINT, Blue must update his beliefs and make a decision (i.e., a choice of option A, B, C, or D in *Decision Making at Two Blue Points*, discussed above). In many cases of real-world importance, the *collections* and *analysis* functions are separated from the operations function, such that the collector and analyst do not know exactly what *decisions* their intelligence will be used to support. Indeed even within the intelligence function itself, there may be a separation between collection and analysis such that the collector does not know exactly what *inferences* his intelligence (e.g., SIGINT) will be used to support. Thus there are several possible solutions to the foraging mission posed by TACTICS, two of which are derived below.

4.3.1 Maximizing the Information Gain at Independent Points

To begin, assume there are two *collections* options for Blue: option 1 is to get SIGINT at point 1, and option 2 is to get SIGINT at point 2. The expected informatic utilities (i.e., expected information gains) are denoted E_1 and E_2, respectively. The collections problem is to compute E_1 and E_2, so that Blue can then select the option (1 or 2) with highest E, i.e., $\max(E_1, E_2)$.

At a given Blue point (1 or 2), the computation of E requires two forms of input. One input is the current probability of attack P_a, i.e., "prior" to receiving SIGINT (which will be received only if this Blue point is chosen). The other input is knowledge of SIGINT reliability, in the form of a "hit rate" (h), "miss rate" (1-h), "false alarm rate" (f), and "correct rejection rate" (1-f). As outlined in *Description* (Section 2), the likelihoods of signals (S = Y or S = N) given ground truth (y or n) are as follows:

$$h = p(Y|y) = 0.60$$

$$1-h = p(N|y) = 0.40$$

$$f = p(Y|n) = 0.20$$

$$1-f = p(N|n) = 0.80.$$

Using u to denote the informatic utility from each possible SIGINT return (Y or N), the expected information gain for SIGINT at a Blue point is given as follows:

$$E = p(Y) * u(Y) + p(N) * u(N).$$

The marginal probabilities p(Y) and p(N) of signals (Y and N) are each computed as the sum of joint probabilities, as follows:

$$p(Y) = p(y) * p(Y|y) + p(n) * p(Y|n) = p * h + (1-p) * f$$

$$p(N) = p(y) * p(N|y) + p(n) * p(N|n) = p * (1-h) + (1-p) * (1-f)$$

where $p = p(y) = P_a$ is the "prior" (before SIGINT) probability of Red attack at the Blue point, and $p(n) = 1-p(y) = 1-p$.

The informatic utilities u(Y) and u(N) depend on the probability of attack before and after SIGINT. More specifically, the gain in information (Shannon & Weaver, 1949) is computed as the KL-divergence (Kullback & Leibler, 1951) of a posterior (after SIGINT) probability distribution P' = {p', 1-p'} relative to a prior (before SIGINT) probability distribution P = {p, 1-p}, where the posterior P'(Y) is computed assuming a signal Y and the posterior P'(N) is computed assuming a signal N. These KL-divergences of P' from P are computed as follows:

$$u(Y) = -\Sigma [P * \log_2 P_Y'] + \Sigma [P * \log_2 P]$$

$$u(N) = -\Sigma [P * \log_2 P_N'] + \Sigma [P * \log_2 P]$$

where each sum is taken over the two probabilities in each distribution, e.g., P = {p, 1-p} = {p(y), 1-p(y)}, P_Y' = {p'(y|Y), 1-p'(y|Y)}, and P_N' = {p'(y|N), 1-p'(y|N)}.

The posterior distributions, P_Y' = {p'(y|Y), 1-p'(y|Y)} and P_N' = {p'(y|N), 1-p'(y|N)}, are computed from the prior distribution P = {p(y), 1-p(y)} = {p, 1-p} and parameters (h, f) of SIGINT, via the application of Bayes Rule as follows:

$$p'(y|Y) = (p * h) / [(p * h) + (1-p) * f)]$$

$$p'(y|N) = [p * (1-h)] / [(p * (1-h) + (1-p) * (1-f)].$$

Thus to recap: The expected information gain E for SIGINT at a Blue point is obtained in four steps:

First compute the marginal probabilities p(Y) and p(N) of each signal, using the prior probabilities P = {p, 1-p} and reliabilities (h, f) of SIGINT.

Then compute the posterior probabilities P' = {p', 1-p'} conditional on each signal (Y and N), via Bayes Rule using the prior probabilities P = {p, 1-p} and likelihoods (reliabilities) of SIGINT.

Then compute the informatic utilities u(Y) and u(N), as the KL-divergences of posterior probabilities P' from prior probabilities P for each signal.

Finally, compute expected utility E as the product of probability * utility summed over both possible SIGINT returns (Y, N).

Figure 9 shows the results for E as a function of P_a = p, assuming h = 0.6 and f = 0.2. This figure shows that E is high when p is small or large. For intermediate values of p, E is low and not very sensitive to p. Thus, if there are no further constraints on P_a at the two Blue points (i.e., if P_{a1} and P_{a2} are independent), then the optimal choice (of point 1 or point 2, to receive SIGINT) will depend (per Figure 9) on the relative magnitudes of P_{a1} and P_{a2}. If P_{a1} is small or large and P_{a2} is intermediate, then the optimal choice is point 1. Likewise, if P_{a2} is small or large and P_{a1} is intermediate, then the optimal choice is point 2. Otherwise the optimal choice depends on the precise values of P_{a1} and P_{a2}.

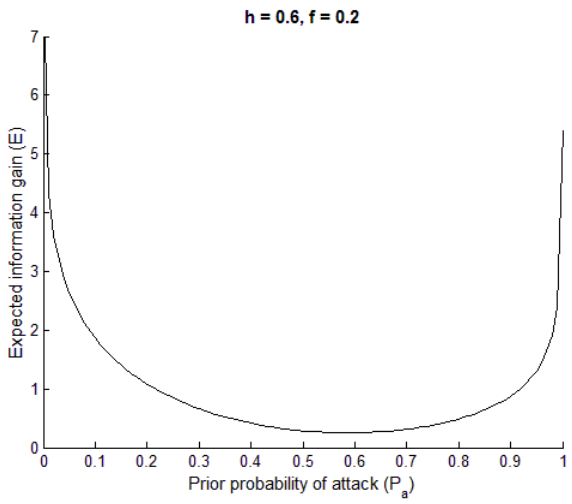

Figure 9: Expected gain in information (E) from SIGINT, as a function of prior probability, with SIGINT hit rate (h) = 0.6 and false alarm rate (f) = 0.2.

4.3.2 Maximizing the Information Gain at Dependent Points

The above analysis applies only if the two or more Blue points are treated as independent. However, in real-world situations, there is often further knowledge that constrains analytical inferences and hence affects the optimal choice for collection. The same is true in TACTICS, where **the Blue analyst knows that Red can attack only one (or neither) Blue point** on a given trial, i.e., $P_{a1} + P_{a2} \leq 1$. With this knowledge, the value of P_{a1} constrains the value of P_{a2}, and vice versa.

To account for this constraint requires a more complex treatment than the previous analysis performed for one point at a time. More specifically, we can define a frame of discernment (set of hypotheses) to cover the set of Red attack possibilities: {A, B, C}, where A = [a_1,~a_2], B = [~a_1,a_2], and C = [~a_1,~a_2]. The corresponding set of probabilities {P(A), P(B), P(C)} = {P_{a1}, P_{a2}, 1-P_{a1}-P_{a2}} is hereafter denoted as the prior probability distribution P = {p_A, p_B, p_C}.

With this prior distribution and SIGINT parameters (h, f), the expected information gain for SIGINT at each Blue point (1 and 2) is as follows:

$$E_1 = p_1(Y) * u_1(Y) + p_1(N) * u_1(N)$$
$$E_2 = p_2(Y) * u_2(Y) + p_2(N) * u_2(N).$$

The marginal probabilities are computed as the sum of joint probabilities, as follows:

$$p_1(Y) = p_1(y) * p_1(Y|y) + p_1(n) * p_1(Y|n) = p_A * h + (1-p_A) * f$$
$$p_2(Y) = p_2(y) * p_2(Y|y) + p_2(n) * p_2(Y|n) = p_B * h + (1-p_B) * f$$

$p_1(N) = p_1(y) * p_1(N|y) + p_1(n) * p_1(N|n) = p_A * (1-h) + (1-p_A) * (1-f)$
$p_2(N) = p_2(y) * p_2(N|y) + p_2(n) * p_2(N|n) = p_B * (1-h) + (1-p_B) * (1-f)$

The informatic utilities, computed as KL-divergences of P' from P are as follows:

$u_1(Y) = -\Sigma [P * \log_2 P_{Y1}'] + \Sigma [P * \log_2 P]$
$u_2(Y) = -\Sigma [P * \log_2 P_{Y2}'] + \Sigma [P * \log_2 P]$

$u_1(N) = -\Sigma [P * \log_2 P_{N1}'] + \Sigma [P * \log_2 P]$
$u_2(N) = -\Sigma [P * \log_2 P_{N2}'] + \Sigma [P * \log_2 P]$

where each sum is taken over the three probabilities in each distribution, e.g., $P = \{p_A, p_B, p_C) = \{P(A), P(B), p(C)\}$, $P_{Y1}' = \{p(A|Y_1), p(B|Y_1), p(C|Y_1)\}$, etc.

The posterior distributions are computed from the prior distributions and likelihoods (reliabilities) of SIGINT, via the application of Bayes Rule. The likelihoods (L) of SIGINT are as follows:

$L_{Y1} = \{P(Y_1|A), P(Y_1|B), P(Y_1|C)\} = \{h, f, f\}$
$L_{Y2} = \{P(Y_2|A), P(Y_2|B), P(Y_2|C)\} = \{f, h, f\}$

$L_{N1} = \{P(N_1|A), P(N_1|B), P(N_1|C)\} = \{1-h, 1-f, 1-f\}$
$L_{N2} = \{P(N_2|A), P(N_2|B), P(N_2|C)\} = \{1-f, 1-h, 1-f\}$.

For example, referring to the likelihood distribution L_{Y1}, $P(Y_1|A)$ refers to the probability of receiving a signal Y at point 1 assuming Red option A (i.e., Red attack at point 1). This is the hit rate, h. Conversely, $P(Y_1|B)$ refers to the probability of receiving a signal Y at point 1 assuming Red option B (i.e., Red attack at point 2, which means no Red attack at point 1). This is the false alarm rate, f. Similarly, $P(Y_1|C)$ refers to the probability of receiving a signal Y at point 1 assuming Red option C (i.e., no Red attack at point 1 or point 2, which means no Red attack at point 1). This is also the false alarm rate, f. The likelihood distribution L_{Y2} is obtained by the same logic.

Referring to the likelihood distribution L_{N1}, $P(N_1|A)$ refers to the probability of receiving a signal N at point 1 assuming Red option A (i.e., Red attack at point 1). This is the miss rate, 1-h. Conversely, $P(N_1|B)$ refers to the probability of receiving a signal N at point 1 assuming Red option B (i.e., Red attack at point 2, which means no Red attack at point 1). This is the correct rejection rate, 1-f. Similarly, $P(N_1|C)$ refers to the probability of receiving a signal N at point 1 assuming Red option C (i.e., no Red attack at point 1 or point 2, which means no Red attack at point 1). This is also the correct rejection rate, 1-f. The likelihood distribution L_{N2} is obtained by the same logic.

Thus to recap: The expected information gains E_1 and E_2 for SIGINT at Blue points 1 and 2, subject to the constraint that Red can attack at only one (or neither) point, are obtained in four steps:

First compute the marginal probabilities of each signal at each point, using the prior probabilities and reliabilities (h, f) of SIGINT.

Then compute the posterior probabilities conditional on each signal at each point, via Bayes Rule using the prior probabilities and likelihoods (reliabilities) of SIGINT.

Then compute the informatic utilities, as the KL-divergences of posterior probabilities from prior probabilities for each signal at each point.

Finally, compute expected utility as the product of probability * utility summed over both possible SIGINT returns at each point.

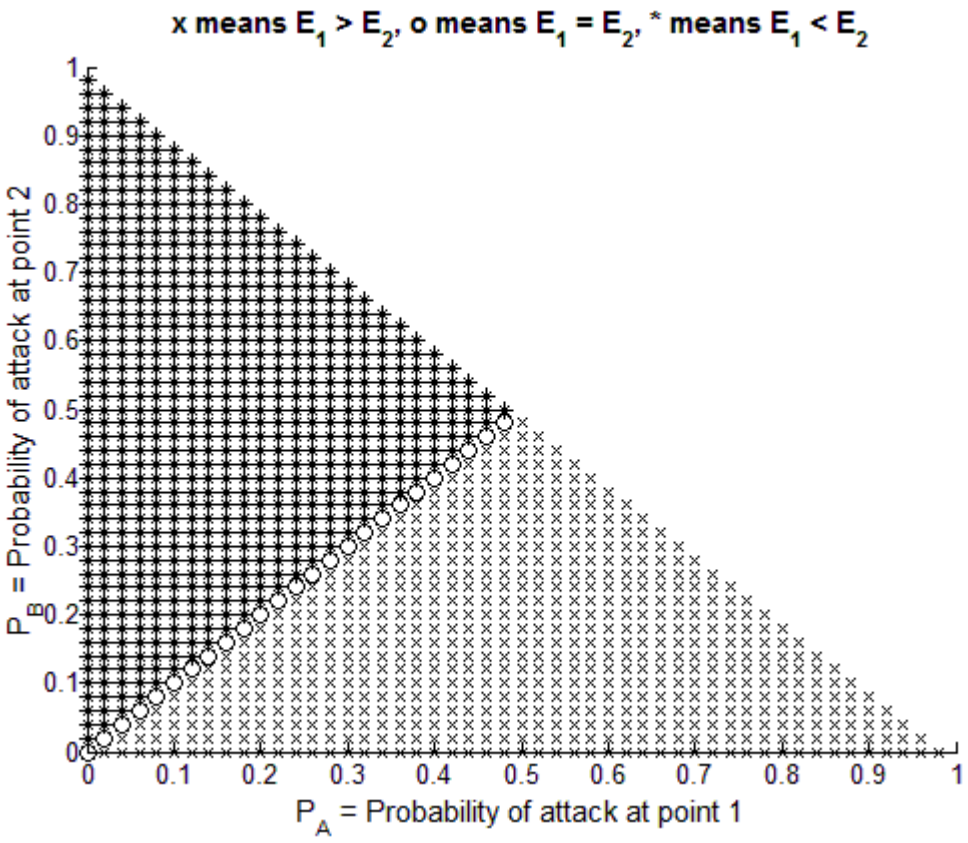

Figure 10: Difference in expected information gains (E_1 - E_2) for SIGINT at two points (1 and 2). Each point has a different "prior" (before SIGINT) probability of attack, P_A at point 1 and P_B at point 2. SIGINT reliabilities are h = 0.6 and f = 0.2. Refer to text for further details.

Figure 10 shows the results assuming SIGINT reliabilities are h = 0.6 and f = 0.2. This figure plots the difference in expected information gains, $E_1 - E_2$, as a function of P_A (prior probability at point 1) and P_B (prior probability at point 2). The figure shows that the difference $E_1 - E_2$ is > 0 (denoted by the symbol x) whenever $P_A > P_B$, and the difference $E_1 - E_2$ is < 0 (denoted by the symbol *) whenever $P_A < P_B$. In other words, the optimal point (1 or 2) at which to request SIGINT is whichever point has the higher prior probability (i.e., P_A for point 1 or P_B for point 2).

This result is consistent with a ***confirmation preference***, aka "positive test strategy", which is known to be an optimal strategy for seeking information in hypothesis testing of many realistic situations (Klayman & Ha, 1987). The result is also consistent with the normative solution computed for the Phase 1 challenge problem (Burns, Greenwald, & Fine, 2014), which was to seek SIGINT on the enemy group with the highest attack probability. Importantly, this *confirmation preference* is not a ***Confirmation Bias*** per se because the strategy is actually optimal (i.e., not sub-optimal).

4.4 Inferencing (Forensic)

In Missions 2, 4, and 5, forensic analyses are required to infer the Red style from previous attacks. These forensic inferences are needed to support prognostic inferences of the Red attack probability P_p, as a function of P and U, on each trial. A normative (Bayesian) solution can be computed assuming there is no change in Red style over time. This solution applies rigorously to Mission 2 but only approximately to Missions 4 and 5 (where there is a change in Red style).

The solution, which assumes no change in Red style over time, is computed as follows: First, actual data from SIGACTS (attack or ~attack) on all previous trials are used to compute the total frequency (F) of attack, i.e., the number of attacks (n) divided by the number of trials (t): F = n/t. Then, the likelihood (probability) of actually observing this frequency (F) is computed for two generative models of attack frequency (f). For example, in Mission 4 these two generative models would be $f_{Passive}$ and $f_{Aggressive}$, which are computed from the BLUEBOOK values of Red attack probability – with each BLUBOOK value (corresponding to a P,U combination) weighted by the actual frequency of the associated P,U conditions. Finally, the likelihood of observing F for each generative model f can be computed from the binomial distribution, which gives the probability $p(F|f) = [t! / (n! (t-n)!)] * f^n * (1-f)^{(t-n)}$. Assuming a uniform prior distribution in which each generative model is equally likely, $p(f_{Passive}) = p(f_{Aggressive})$, the posterior probability p(f|F) of each Red style is computed from Bayes Rule to obtain p(Passive) and p(Aggressive).

Given this forensic inference of p(Passive) and p(Aggressive), values of P and U (from OSINT and IMINT) can be used along with BLUEBOOK values of Red attack probability to compute: P(Attack| IMINT, OSINT) = p(Passive) * P(Attack| BLUEBOOK(Passive), IMINT, OSINT) + p(Aggressive) * P(Attack| BLUEBOOK(Aggressive), IMINT, OSINT).

Note that for Mission 2, the calculation of generative model frequency f would be based on $P_c * P_p$, where P_c is given by HUMINT for each trial (but P_c is the same for each model f). In Missions 4-5, $P_c = 1$ always. For Mission 5, the calculation of p(F|f) would be performed separately for each P,U cell of the BLUEBOOK, and then cells for each style (P-sensitive and U-sensitive) would be aggregated to obtain p(P-sensitive) and p(U-sensitive).

5 Evaluation

This section outlines the methods to be employed in Comparative Performance Assessment, Cognitive Fidelity Assessment, and Neural Fidelity Assessment. The T&E approach for each type of assessment is similar to that of Phase 1 (Burns, Greenwald, & Fine, 2014), therefore this section focuses on differences in Phase 2.

5.1 Comparative Performance Assessment (CPA)

Comparative Performance Assessment (CPA) will assess a model's success in matching human performance, per the BAA Table 4 criterion of a *65% success rate* (for Phase 2). The primary data represent *judgments* in the form of probability distributions, reported by humans and models on stages of trials of missions that require *inferencing* (see *Variations,* Section 3). These data on judgments are assessed by an ***Absolute Success Rate*** (ASR), discussed in Section 5.1.1 below. Additional data represent *choices* made on stages of trials of missions that require *decision-making* and *foraging* (see *Variations,* Section 3). These data on choices are assessed by a ***Relative Match Rate*** (RMR), discussed in Section 5.1.3 below.

For both ASR and RMR, human data from individual participants are aggregated into measures of *average performance* in order to assess neural models. This is discussed further in Section 5.1.2 below. Also, ASR and RMR are subject to *weighting factors* that are applied to each mission in computing a model's overall performance on CPA. This is discussed further in Section 5.1.4 below. The methods and missions for CPA are summarized in Table 3.

Table 3: Methods and missions for Comparative Performance Assessment (CPA).

Process	Method	Mission				
		1	**2**	**3**	**4**	**5**
Inferencing	ASR	X	X	X	X	X
Decision-making	RMR		X	X	X	X
Foraging	RMR			X		

5.1.1 Absolute Success Rate (ASR)

The primary measures of sensemaking are probability distributions reported by humans and models on stages of trials of missions. In Phase 1, a model distribution was compared to the human distribution using a ***Relative Success Rate*** (RSR) that accounts for two forms of similarity. One similarity is between the human distribution P and a model distribution M, denoted S_{PM}. The other similarity is between the human distribution P and a "random" (maximum entropy) distribution R, denoted S_{PR}. These similarity measures, in turn, are based on an information-theoretic (Shannon & Weaver, 1949) measure of "divergence" (Kullback & Leibler, 1951), denoted K, between two probability distributions.

All of these quantities (RSR, S_{PM}, S_{PR}, K_{PM}, and K_{PR}) are defined and discussed in the Phase 1 Challenge Problem Design and Test Specification (Burns, Greenwald, & Fine, 2014). For convenience the equations are repeated here as follows:

$$K_{PM} = -\Sigma [P * \log_2 M] + \Sigma [P * \log_2 P]$$
$$K_{PR} = -\Sigma [P * \log_2 R] + \Sigma [P * \log_2 P]$$

$$S_{PM} = 100\% * (2 \wedge -K_{PM})$$
$$S_{PR} = 100\% * (2 \wedge -K_{PR})$$

$$RSR = \max[0\%, (S_{PM} - S_{PR}) / (100\% - S_{PR})]$$

where P, M, and R are discrete probability distributions, e.g., P = {P, 1-P} for the case of two hypotheses; P is the human distribution; M is a model distribution; and R is the "random" (uniform) distribution, e.g., R = {0.50, 0.50}.

Using these equations, the RSR for one data point (i.e., a probability distribution reported on a stage of a trial of a mission) is computed as follows: First, K_{PM} and K_{PR} are computed from P, M, and R. These K values range from 0 (perfect match of model to human) to infinity (worst possible match of model to human). Then, the K values are converted to S values that range from 0% (worst match, K is infinite) to 100% (perfect match, K is zero). Finally, S_{PM} is scaled by S_{PR} and the final RSR is limited to values $0\% \leq RSR \leq 100\%$.

The scaling of S_{PM} by S_{PR} is performed because even a poor match of model to human will often produce $K_{PM} < 1$ and hence $S_{PM} > 50\%$. Per RSR, a model's match to human data is therefore measured on a scale of 0-100% *relative* to a random model's match to human data. If a neural model matches human data worse than the random model, then RSR is set to its minimum value of 0%. Otherwise RSR > 0%. For example, if M matches P with similarity $S_{PM} = 80\%$, and R matches P with similarity $S_{PR} = 40\%$, then M would score (80 – 40) / (100 – 40) = 67%.

The above approach from Phase 1 is problematic for Phase 2, because in Phase 2 it is more difficult to design trials for which human performance is far from random (e.g., a uniform probability distribution {P, 1-P} in which P = 1-P). In that case, there is little or no potential for any model to outperform a random model, so the "relative" success measured by RSR is near zero even when the "absolute" difference between model and human distributions is small.

To address this issue, Phase 2 will adopt a different metric for use in CPA. The new metric is an ***Absolute Success Rate (ASR)***, defined as follows:

$$ASR = \max[0\%, (100\% - 2 * RMS_{PM})]$$

where RMS_{PM} is the Root Mean Squared error between the human (P) and model (M) distributions.

For example, assume the human distribution is P = {70%, 30%} and the model distribution is M = {53%, 47%}. In that case RMS_{PM} = 17% and ASR = 66%. Thus with two hypotheses, the Phase 2 criterion of 65% would be satisfied by a model with $RMS_{PM} \leq 17.5\%$.

Besides ASR, which will be used to score model performance, RSR will also be computed as an indication of how much predictive capability a model has relative to (i.e., over and above) a random solution.

Note that the definition of ASR above includes a factor of two. This factor is derived from a principled approach to address the fact that RMS errors are dependent on the number of hypotheses. The factor of two assumes there are two hypotheses, as there are on all trials of Phase 2, e.g., {attack, ~attack}. The factor accounts for the difference between a maximum entropy distribution and minimum entropy distribution when there are two hypotheses. That is, the RMS distance between maximum entropy {50%, 50%} and minimum entropy {100%, 0%} is 50%, so the "zero-value" of ASR is set to occur when RMS = 50%, such that the factor is 100%/50% = 2. By the same logic, with four hypotheses the RMS distance between maximum entropy {25%, 25%, 25%, 25%} and minimum entropy {100%, 0%, 0%, 0%} is 43.3%, so the appropriate ASR factor would be 100%/43.3% = 2.31 (rather than 2).

Per the above logic, ASR is scaled by the difference between a maximum-entropy (random) and minimum-entropy distribution, in order to account for the number of hypotheses in probability distributions. In that sense there is some notion of "relative" scaling. But this is much different from the "relative" performance that is modeled by RSR, because ASR can be high even when the human distribution is nearly random. Therefore ASR is indeed an ***Absolute Success Rate*** that differs markedly from the *Relative Success Rate* RSR.

5.1.2 Average Performance

As discussed above, ASR is concerned with *judgments* reported in the form of probability distributions. In that case, the average human performance at one data point (i.e., a stage of a trial of a mission) is an *average probability distribution* – computed as a simple average across the N human subjects. On the other hand, RMR (discussed below) is concerned with *choices* reported in decision-making and foraging, where each human subject makes a forced choice among options (e.g., option A or option B). In that case the average human performance at one data point is an *aggregate frequency distribution* – computed by summing the number of responses for each option and dividing by the number of human subjects.

Per the BAA, CPA reduces individual human responses to *average human performance* in order to assess model predictions. T&E requires that a model compute a comparable *average model performance*. It is the responsibility of the modeler (not T&E) to determine how the average model performance is computed. It is also the responsibility of the modeler's software to compute average model performance and report each data point as a single response (i.e., not a collection of individual model responses).

5.1.3 Relative Match Rate (RMR)

Absolute Success Rate (ASR), discussed in Section 5.1.1, applies to human *judgments* that are reduced to *average probability distributions*. A different metric, called ***Relative Match Rate (RMR)***, applies to human *choices* that are reduced to *aggregate frequency distributions*. Mathematically, these frequency distributions, e.g. {A%, B%}, are equivalent to discrete probability distributions in that each value is a number 0-100% and the numbers sum to 100%. However, the single (forced choice) response of a model on each trial is akin to a frequency of {100%, 0%} or {0%, 100%}. Thus RMR differs from RSR in computing the relative match of the model's forced choice responses to human forced choice frequency distributions.

The calculation of RMR on each trial is performed much like in Phase 1. First, the option with highest frequency in the average human data is identified as f_{max}. Second, the human frequency corresponding to the model choice is identified as f_{mod}. Finally, the ratio f_{mod}/f_{max} is taken as the measure of RMR on the trial.

For example, assume the average human frequencies for options {A, B} on a trial are {60%, 40%}. A model that chooses option A would score 60/60 = 100%, and a model that chooses option B would score 40/60 = 67%. By this method, a model scores 100% for a choice that matches the dominant human response. The model scores a ratio amount (< 100%) for a choice that does not match the dominant human response, and the ratio decreases as the non-dominant human frequency decreases relative to the dominant human frequency.

This approach applies to any choice between two options, e.g., Blue *decision-making* choices between {d, ~d} in Mission 2, or Blue *forging* choices between {point 1, point 2} in Mission 3. The same approach would extend to larger sets involving three, four, or more options.

5.1.4 Relative Weighting

As described above, ASR or RMR will be computed for each data point (stage on trial) in one or more missions, see Table 3. Within a mission, all *judgment* data points will be weighed equally in computing an average ASR for the mission, and all *forced choice* data points will be weighed equally in computing an average RMR for the mission. Similarly, all missions will be weighed equally in computing the overall ASR and overall RMR. Finally, ASR and RMR will be weighed equally in computing the overall score of a model on CPA.

5.2 Cognitive Fidelity Assessment (CFA)

Cognitive Fidelity Assessment (CFA), like Comparative Performance Assessment (CPA), is concerned with how well a model predicts human performance – but more specifically with a focus on ***cognitive biases***. The two assessments are clearly related, because any model that closely matches human data per CPA will naturally replicate behavioral biases. However, CFA is distinguished by an explicit focus on cognitive biases, to encourage generalization and application of models and insights to real-world intelligence and operations. Per BAA Table 4, for Phase 2, a model is required to exhibit 5 of the 8 biases listed in BAA Table 3.

CFA requires formal (computational) definitions of biases – i.e., so the existence of bias in human data can be identified in experiments, and so the extent of such bias exhibited by neural models can be evaluated (and possibly extrapolated in *Transition*, see Section 6). These definitions, in turn, require a reference model or "benchmark" from which biases can be measured objectively. Although omniscient benchmarks like "ground truth" or "hindsight" might be chosen, these are unfair standards because they assume more information that the sensemaker himself has when he needs to make sense. Thus the proper standard is a *normative* model (Edwards, 1954; Edwards, 1961; Edwards, et al., 1963), which is given the same information (knowledge and data) as the human sensemaker but computes ***Bayesian*** judgments (in *inferencing*) and choices (in *decision-making* and *foraging*).

By this approach, normative solutions provide a critical foundation for defining and measuring cognitive biases. The necessary *Solutions* are derived in Section 4, as the first step in preparing for CFA. The next step is to describe and define the BAA (Table 3) biases, relative to these normative solutions (or relative to some other benchmarks when normative solutions are intractable). The last step in preparing for CFA is pilot testing of human subjects, in order to establish at which stages of which trials of which missions the humans exhibit biases per the definitions.

All eight BAA biases were described briefly in *Variations* (Section 3), as a preview of how various missions might elicit these biases. The following sections provide more detailed descriptions and computational definitions, with each section focusing on biases for one of the cognitive processes outlined in *Variations* (Section 3), namely: *inferencing* (Section 5.2.2), *decision-making* (Section 5.2.3), and *foraging* (Section 5.2.4). But before addressing the biases individually, it is useful to consider them collectively, and especially to highlight the difference between ***heuristics*** and ***biases***.

5.2.1 Heuristics and Biases

As defined in the literature on judgment and decision-making (Kahneman, et al. 1982; Gilovich, et al., 2002), *heuristics* are simplified processes (aka "rules of thumb") in human thinking that cause subjective judgments and decisions to deviate from normative (optimal) judgments and decisions. The deviations themselves, measured objectively, are called *biases*. For example, a heuristic known as ***Representativeness*** may produce a bias known as ***Change Blindness***; a heuristic known as ***Availability*** may produce a bias known as ***Satisfaction of Search***; and a heuristic known as ***Anchoring and Adjustment*** may produce a bias known as ***Confirmation Bias***.

The difference between a "heuristic" and a "bias" is important for three reasons. First, the BAA includes the six heuristics and biases noted above but refers to them all as "biases". Because half of them are actually heuristics, the BAA biases may be somewhat redundant with respect to the human behaviors that are implied. In CFA, T&E must define distinct behaviors for each of the eight BAA biases – even those that are actually heuristics. Second, only biases are measurable directly from human behavioral experiments, because the associated heuristics are merely conjectures about the cognitive processes that produce biases. This requires that T&E itself make subjective judgments about which heuristics are causing which biases, in order to assess all eight

of the BAA biases (which actually include some heuristics). Finally, there is overlap even among the biases themselves, because different heuristics may produce the same or similar behavioral bias. For example, a ***Persistence of Discredited Evidence*** and a ***Confirmation Bias*** can refer to the same response in which the weighing of evidence is skewed toward "confirming" a favored hypothesis more than it should be by "discredited" evidence.

In theory, a single heuristic may produce different and perhaps even opposite behaviors (biases) in different situations. Similarly, a single bias may actually refer to several different behaviors that stem from different heuristics in different missions of TACTICS. For example, ***Confirmation Bias*** is a broad term (Nickerson, 1998) that can refer to bias in aggregating likelihoods (in inferencing) and/or bias in selecting evidence (in foraging). These issues have been carefully considered in the definition of biases and specification of metrics for CFA, in order to meet the intent as well the content of the BAA's guidance (Table 3 and Appendix F). In so doing, each BAA "bias" will be assigned a formal metric that can be measured directly in human data from the Phase 2 experiment. These metrics are similar to those defined and employed in Phase 1, which included four of the eight biases for Phase 2.

The metrics and missions for CFA are summarized in Table 4. Note that in some cases the metric is the same for different biases, e.g., $N_P < N_Q$. In that case the measured bias is the same, but the postulated heuristic that causes the bias in a context (i.e., stage of trial of mission) is different and consistent with the BAA "bias".

Table 4: Metrics and missions for Cognitive Fidelity Assessment (CFA).

BAA Bias	Metric	Mission				
		1	2	3	4	5
Anchoring and Adjustment	$N_P < N_Q$	X	X			
Persistence of Discredited Evidence	$N_P < N_Q$				X	X
Representativeness	$P > Q$	X	X	X		
Availability	$N_P < N_Q$	X				
Probability Matching	n		X		X	X
Confirmation Bias	f			X		
Satisfaction of Search	s				X	X
Change Blindness	b				X	X

5.2.2 Inferencing

As discussed in *Variations* (Section 3), the main BAA biases associated with *inferencing* are ***Anchoring and Adjustment***, ***Persistence of Discredited Evidence***, ***Representativeness***, and ***Availability***.

One bias can be measured simply by comparing the human probability (P) to the Bayesian probability (Q). That is, in *prognostic inferencing* at the start of a trial in Missions 1-3, it appears from pilot data that humans are typically computing $P_{p,c}$ as the arithmetic average of P_c and P_p. This average is greater than the normative solution given by $Q_{p,c} = P_c * P_p$. Therefore, the bias is measured by $P_{p,c} > Q_{p,c}$. The underlying heuristic is one of ***Representativeness*** in which *capability* and *propensity* are treated as equally *representative* of the composite activity (attack), such that P_c and P_p are averaged to obtain $P_{p,c}$.

For the remaining three inferencing biases, it is useful to distinguish "conservative" from "non-conservative biases – where conservatism is computed by a quantity referred to as *Negentropy* (also used in Phase 1, see Burns, Greenwald, & Fine, 2014). Negentropy ranges from 0% to 100% as entropy ranges from maximum entropy to minimum entropy, and entropy itself refers to the uncertainty across a set of hypotheses. For example, {50%, 50%} represents maximum entropy (0% Negentropy), and {100%, 0%} represents minimum entropy (100% Negentropy). Mathematically, entropy is computed as follows:

$$E_P = -\Sigma\ P * \log_2 P$$

and Negentropy is computed as follows:

$$N_P = (E_{max} - E) / E_{max}$$

where E_{max} depends on the number of hypotheses in the frame of discernment, i.e., $E_{max} = 1$ for the case of two hypotheses, and $E_{max} = 2$ for the case of four hypotheses.

A conservative bias in inferencing is defined as one in which a human extracts less overall certainty than he or she should from the evidence he or she is given (Edwards, 1982), i.e., the distribution P is too "flat". A non-conservative (confirmation) bias in inferencing is the opposite case in which a human assigns too much certainty, i.e., the distribution P is too "peaked". Mathematically, the difference is captured by comparing Negentropy N_P of the human distribution P to Negentropy N_Q of the Bayesian distribution Q. A conservative bias implies $N_P < N_Q$, and a non-conservative bias implies $N_P > N_Q$. Thus, N allows us to distinguish one class of inferencing biases from the opposite class of inferencing biases.

In the case of *forensic inferencing*, in Missions 4 and 5, pilot data suggest that humans are conservative in their estimate of P(style), where the styles are: Passive and Aggressive in Mission 4; P-sensitive and U-sensitive in Mission 5. Although we only compute a quasi-Bayesian solution, under the assumption that there is no change in Red style during these missions (see Section 4.4), pilot data show that humans are more conservative than this quasi-Bayesian – especially after the change in Red style. This conservatism ($N_P < N_Q$) can be

characterized as ***Persistence of Discredited Evidence***, because too much uncertainty (conservatism) "persists" in the human distribution even though early evidence (from SIGACTS) has been "discredited" by later evidence (from SIGACTS).

Returning to the task of *prognostic inferencing*, in Missions 1-2, two additional conservative biases can be measured by $N_P < N_Q$. First, in Mission 1, pilot data show that humans are conservative in reporting the distribution $\{P_t, 1\text{-}P_t\}$, which represents the probability of {attack ~attack} based only on SIGINT. In effect, humans are failing to compute a Bayesian-normalized posterior and instead report the raw SIGINT likelihoods (see Section 4.1). This is attributed to ***Availability*** as the SIGINT likelihoods are readily available whereas the normative probabilities $\{P_t, 1\text{-}P_t\}$ require further computation (i.e., normalization over the hypotheses {attack, ~attack}). The bias is measured only in Mission 1 because this is the only mission for which subjects are required to report P_t.

The final conservative bias occurs in Bayesian updating of $P_{p,c}$ with P_t to compute $P_{t,p,c}$. Pilot data suggest that humans are once again averaging, much like in ***Representativeness*** discussed above. However, here the normative solution is to compute a Bayesian-normalized product of $P_{p,c}$ and P_t, rather than a simple product. For this Bayesian update, the conservative bias stemming from averaging is characterized as ***Anchoring and Adjustment*** – because there are effectively two "anchors" ($P_{p,c}$ and P_t) and the inadequate adjustment is to compute an arithmetic average of the anchors rather than a Bayesian-normalized product. Like the other conservative biases mentioned above, this ***Anchoring and Adjustment*** is measured by $N_P < N_Q$.

5.2.3 Decision-Making

As discussed in *Variations* (Section 3), the main BAA bias associated with *decision-making* is ***Probability Matching***. In CFA this bias is assessed for Missions 2, 4, and 5. On these missions, Blue decisions to divert (d) or not divert (~d) represent choices that will be assessed using the metric RMR in CPA. Thus CFA uses a different measure of performance, relative to normative solutions (not considered in CPA), aimed specifically at the bias of ***Probability Matching***.

In particular, on each trial the normative *Solutions* (Section 4) can be used to compute the optimal (Bayesian) Blue choice. We expect human choices will sometimes deviate from the Bayesian choices, for various reasons. For example, humans may be biased in their estimation of expected utility for each option, E_d and $E_{\sim d}$. On the other hand, humans may properly compute expected utilities (or at least their relative magnitudes as needed to make optimal choices, i.e., $E_d > E_{\sim d}$ or $E_d < E_{\sim d}$) but ***sometimes not choose*** the option (d or ~d) with higher expected utility. That behavior would imply ***Probability Matching***, where humans are presumably choosing the two options at frequencies governed by their relative expected utilities as scaled by a multinomial logit function (see Burns & Demaree, 2009).

In Mission 2 (and other missions), humans are not asked to report expected utilities. Therefore any bias in decisions would include bias in estimating expected utilities ***and*** bias in applying the estimates per ***Probability Matching***. Nevertheless, T&E will compute the deviation in decisions (relative to normative solutions) and use those errors as a measure of the BAA bias for ***Probability Matching***. On each trial, a number 1 or 0 will be assigned to a human's decision. The

number 1 means the human chose the normative option (d or ~d), and the number 0 means he or she did not. Across all subjects, the numbers (1 or 0) will be used to compute an average number n on each trial of each mission. Finally, the average number n_H across all trials of the mission will be taken as the measure of ***Probability Matching***. For example, if $n_H = 1$ then there is no bias in decisions relative to normative solutions. As n_H decreases there is more bias, and at least some (perhaps much) of this bias might be attributed to the mechanisms of probability matching.

A similar calculation will be done for a model, to compute an equivalent average number n_M across all trials of a mission. This model number n_M will be compared to the human number n_H in order to assess ***Probability Matching***. The comparison of n_M to n_H will assessed by a ***Marginal Success Rate*** (MSR, discussed in Section 5.2.6 below).

With respect to *Variations* (Section 3), some of the *inferencing* biases may also be exhibited in *decision-making* – especially ***Availability***, ***Representativeness***, and ***Anchoring and Adjustment***. The reason, mentioned above, is that these are actually heuristic processes (not biases) and such heuristics may apply to inferencing, decision-making, or other cognitive processes. For example, a decision-making situation may be *representative* of familiar situations, and/or the outcome of an earlier decision may especially vivid or otherwise *available* from memory, and either or both phenomena may cause a human to be *anchored* to a sub-optimal strategy.

Therefore, in theory these heuristics might be measured in the context of decision-making as well as in the context of inferencing. However, in TACTICS human subjects are making choices that have outcomes, so the sequences of choices and outcomes across trials are different for each subject. This makes it infeasible to assess these heuristics (biases) in the context of decision-making or foraging (Section 5.2.4), so instead they are assessed only in the context of inferencing (Section 5.2.2).

5.2.4 Foraging

The remaining biases listed in Table 4 are assessed in the context of foraging. As discussed in *Variations* (Section 3), Mission 3 involves *prognostic foraging* (to obtain SIGINT) whereas Missions 4 and 5 involve *forensic foraging* (to review SIGACTS).

First, for prognostic foraging in Mission 3, the variable P_a is a measure of the humans' confidence in Red attack at each Blue point (1 or 2), i.e., P_{a1} and P_{a2}. In the case of a "pure" confirmation preference, humans would always seek SIGINT on the point (1 or 2) with higher P_a in order to "confirm" their belief. Instead we expect (based on pilot data) that humans will often but not always do so, as measured by a frequency f. Therefore, similar to the numbers n_H and n_M computed for ***Probability Matching***, we will compute numbers f_H and f_M as a means of assessing ***Confirmation Bias***. The comparison of f_M to f_H will be assessed by a ***Marginal Success Rate*** (MSR, discussed in Section 5.2.6 below).

Here it is important to note that, although the term ***Confirmation Bias*** is being used by T&E per BAA, the actual behavior here is a confirmation ***preference*** and it is not a confirmation ***bias*** per se. As found in *Solutions* (Section 4), the optimal choice (under reasonable assumptions for maximizing information gain) is to seek SIGINT on the Blue point (point 1 or point 2) with the

higher P_a. In that sense the only "bias" is actually a conservative (not confirmation) bias in which humans ***do not always*** exhibit the confirmation preference. But even the status of this conservative behavior as a "bias" is not so clear cut, because the normative solution assumes there is no second-order uncertainty (i.e., a probability of the probability P_a). A human being who feels he or she does not know P_a with certainty may adopt a form of ***Probability Matching***, where the frequency at which he or she does not choose the point with highest P_a increases as second-order uncertainty increases. Indeed that very strategy has been shown to be optimal (normative), in the context of other tasks with second-order uncertainty for which humans are found to exhibit ***Probability Matching*** (Burns & Demaree, 2009).

Finally, two additional biases will be assessed in the context of forensic foraging through batch plots in Missions 4 and 5. These two biases, ***Change Blindness*** and ***Satisfaction of Search***, are somewhat different from the other biases in three respects. First, these biases do not typically appear in the literature on judgment and decision making (Kahneman, et al., 1982; Gilovich, et al., 2002) or in discussions of how that literature may apply to the practice of intelligence analysis (Heuer, 1999). Second, it is not clear what assumptions should be made in computing normative solutions for ***Change Blindness*** and ***Satisfaction of Search***.

The literature on these biases implies that *any changes* should be detected and *all searches* should be exhaustive, yet that is clearly infeasible and unreasonable for a person or agent that has limited resources. Moreover, a normative solution that did address such limitations would also need to make assumptions about the potential benefits of detecting changes or completing searches – and these assumptions would be very dependent on the context of the change or search situation. Finally, ***Change Blindness*** (Macknik, et al., 2008) and ***Satisfaction of Search*** (Berbaum, et al., 1990) are largely biases in attention and visual perception, and these lower-level cognitive processes are outside the scope of the ICArUS BAA.

In that light ***Change Blindness*** and ***Satisfaction of Search*** are treated somewhat differently from the other BAA biases, and defined relative to omniscient knowledge and unlimited effort – such that *any* change that is not successfully detected will be characterized as a ***Change Blindness***, and *any* search that is not completed will be characterized as a ***Satisfaction of Search***. In effect, the bias will be defined as a specific change not detected or search not completed. For example, in Missions 4 and 5, if Red tactics actually change on trial t, then the extent of ***Change Blindness*** will be measured by the number b_H of trials it takes for subjects to detect the change (measured by a report of P(style) > 50% for the correct post-change style).

Similarly, when a search though "batch plots" of previous trials is required to detect Red's style, the extent of ***Satisfaction of Search*** will be measured by the fraction s_H of all items (on average across subjects) searched in "mouse clicks" associated with batch plots. Like the numbers n (for ***Probability Matching***), f (for ***Confirmation Bias***), and b (for ***Change Blindness***), the number s (for ***Satisfaction of Search***) will be assessed by comparing the model value s_M to the human value s_H and computing the ***Marginal Success Rate*** (MSR, discussed in Section 5.2.6 below).

5.2.5 Simple Match Rate (SMR)

As discussed in Section 5.2.2, the *inferencing* biases (***Anchoring and Adjustment***, ***Persistence of Discredited Evidence***, ***Representativeness***, and ***Availability***) are all defined by some measure of probability or Negentropy in an inequality (human relative to Bayesian). At each stage of each trial of a mission, the model either satisfies the same inequality as humans and is assigned a score of 1, or the model does not satisfy the same inequality as humans and is assigned a score of 0. The scores are then summed over a mission to obtain a fraction (0-100%), called the ***Simple Match Rate*** (SMR). All missions for which a bias is assessed (see Table 4) will be weighted equally in computing an overall SMR for that bias. The resulting score will be compared to the BAA passing threshold of > 65% (Phase 2) for each bias.

5.2.6 Marginal Success Rate (MSR)

The biases in *decision-making* (***Probability Matching***) and *foraging* (***Confirmation Bias***, ***Satisfaction of Search***, and ***Change Blindness***) are all defined by a single number (i.e., n, f, s, or b) computed for humans (e.g., n_H) and a model (e.g., n_M). Each number applies to a mission, and the number for each bias (on each mission) is assessed by a ***Marginal Success Rate*** (MSR), defined below.

Given a number n_H from humans and a corresponding number n_M for a model, the quantity $|n_H - n_M| / n_H$ provides a proportional measure of error or "failure" of the model. Therefore a measure of success is $1 - (|n_H - n_M| / n_H)$. When $n_M < n_H$, this measure of success is always > 0 and < 1. When $n_H < n_M < 2*n_H$, the measure of success is also > 0 and < 1. However, when $n_M > 2*n_H$ then the measure of success is < 0, so a "floor" is imposed to keep it = 0. The marginal success rate is thus defined as follows:

$$MSR = \max[0, 1 - (|n_H - n_M| / n_H)].$$

For example, assume $n_H = 0.8$. In that case, a model with $n_M = 0.6$ would score MSR = 75%, and a model with $n_M = 1.0$ would also score MSR = 75%. Substituting other symbols for n, the same measure of ***Marginal Success Rate*** (MSR) applies to f, s, and b.

Like SMR above, results for MSR are averaged across missions with equal weighting of each mission on which the bias is assessed (see Table 4). The resulting score will be compared to the BAA passing threshold of > 65% (Phase 2) for each bias.

5.3 Neural Fidelity Assessment (NFA)

CPA and CFA are quantitative assessments, hence sensitive to details of challenge problem design. Neural Fidelity Assessment (NFA) performs qualitative assessments, using methods that would apply to any challenge problem design. Details of the NFA approach and schedule, for Phase 2 as well as Phase 1, have already been documented in the Phase 1 Challenge Problem Design and Test Specification (Burns, Greenwald, & Fine, 2014). Per BAA Table 4, for Phase 2, a model is required to faithfully represent 5 of 7 key brain systems.

6 Transition

As outlined in *Introduction* (Section 1), the Phase 2 challenge problem is intended to serve two purposes. The primary purpose, discussed in earlier sections, is to provide a rigorous test-bed for measuring and modeling human sensemaking performance. The secondary purpose is to aid in relevant ***Transition And Communication To Intelligence Community Stakeholders***. This purpose, like the primary purpose, is accomplished by the computational design of TACTICS – which enables a relational mapping to real-world cases of geospatial intelligence.

6.1 Relational Mapping

The mapping highlights six specific types of intelligence analysis that are modeled by variables of TACTICS, namely: *vulnerability* analysis (P), *opportunity* analysis (U), *capability* analysis (P_c), *activity* analysis (P_t), *frequency* analysis (F_t), and *intentionality* analysis (P_a). All six types of analyses were observed across 26 real-world case studies, developed in Descriptive (Cognitive) Task Analysis (MITRE, 2013), via structured interviews with analysts and reviews of published articles. These case studies informed challenge problem design, and a post-design review was performed to make the mapping explicit.

Results of the review are provided in Table 5, showing the six types of analyses and associated variables of TACTICS for each of the cases by title (MITRE, 2013). The Xs in this table are admittedly subjective judgments and are probably incomplete, as they are based on short stories by which the case studies are documented. Nevertheless, the mapping does suggest that each case study involves at least one of the six types of analysis, and most cases involve two or more of the six types. In making this mapping, the following questions were used to judge if a type of analysis (P, U, P_c, P_t, F_t, or P_a) applied ("yes" = X) or not ("no" = blank) to each case study:

> P: Does the analysis ***model spatial constraints on probabilities of activities***, such as proximity or other properties?
>
> U: Does the analysis ***model spatial constraints on utilities of activities***, such as density or other properties?
>
> P_c: Does the analysis ***model temporal constraints on probabilities of activities***, such as recency or other properties?
>
> P_t: Does the analysis ***exploit current reports on probabilities of activities***, such as signals from SIGINT data?
>
> F_t: Does the analysis ***review previous reports of activities and frequencies***, such as "hot-spot" (heat map) plots of SIGACTS?
>
> P_a: Does the analysis ***involve predictions (prognostic) or explanations (forensic)*** of operations (how) and intentions (why) – i.e., beyond merely observations (who, what, when, and where) and visualizations of activities and frequencies?

Table 5: Mapping variables of TACTICS to case studies of intelligence.

No.	Title of Case Study	P	U	P_c	P_t	F_t	P_a
1	Clinical vs. Actuarial Geospatial Profiling Strategies	X				X	
2	Route Security in Baghdad	X	X			X	X
3	International Security Assistance Force Handoff	X	X	X		X	X
4	Explosively Formed Penetrator Placement	X	X	X		X	X
5	Finding Osama Bin Laden	X	X	X			
6	Geospatial Abduction Problems	X				X	
7	Mapping of Cholera in Nineteenth-Century London					X	
8	Clandestine Airstrips in Guatemala	X					
9	Mapping of Arsenic in Twentieth-Century Bangladesh					X	
10	Complexity and Accuracy of Geospatial Profiling Strategies	X				X	
11	Geospatial Analysis of Terrorist Activities	X	X			X	
12	District Control					X	X
13	Tunisian Refugee Flow			X			
14	Improvised Explosive Device (IED) Use in Afghanistan and Pakistan					X	
15	Gang Roundup					X	
16	Gang Geographic Movement					X	
17	Predicting Mortgage Fraud	X	X			X	X
18	Tracking High-Value Cargo	X	X	X	X		X
19	Environmental Study	X		X			
20	Trench Mystery	X	X	X		X	X
21	IED Attack Patterns	X	X	X		X	X
22	Underground Facility	X	X	X		X	X
23	Memphis Airport Communications Failure	X	X				
24	Banking Infrastructure	X	X				
25	The Lone Reconnaissance Vehicle	X	X	X	X		X
26	Road Network Impact on Insurgency	X	X	X		X	

Table 5 shows that the majority of cases involve *vulnerability* (P) analysis and *frequency* (F_t) analysis. The *vulnerability* (P) analyses typically employ various distance functions by which suitability is modeled, much like spatial proximity of a Blue point to the Blue border constrains the probability P (vulnerability) in TACTICS. The *frequency* (F_t) analyses typically produce "dot plots" of historical activities, overlaid on geographic displays, much like the "batch plots" in TACTICS.

In about half of the 26 cases, there was also *utility* (U) analysis, and/or *capability* (P_c) analysis, and/or *intentionality* (P_a) analysis. Of particular interest are the 10 stories of *intentionality* (P_a) analysis, because these are the cases that most clearly go beyond suitability analysis to require ***sensemaking*** – i.e., in *predictions* and *explanations*, per the definition of sensemaking outlined in *Introduction* (Section 1) and *Definitions* (Section 7). Referring to Table 5, two cases of *predicting* intentionality involve *activity* (P_t) analysis to support the estimation of P_a prognostically, whereas eight cases of *explaining* intentionality involve *frequency* (F_t) analysis to support the estimation of P_a forensically.

As discussed in *Description* (Section 2), TACTICS involves all six types of analyses – although the focus is on intentionality analysis (P_a) as Blue's main task is to *predict* the probability that Red will attack and to *explain* Red tactics – i.e., because these are the key functions of sensemaking. The various other analyses are greatly simplified in TACTICS, compared to real-world intelligence, to the point where results for most individual types of *suitability* (capability, activity, etc.) analyses are computed by the "system" and provided to Blue as INT "data" along with associated ***likelihoods*** (probabilities). This makes the task posed by TACTICS closest to that of an "all-source" analyst who acquires and exploits data from various geospatial intelligence sources (OSINT, IMINT, HUMINT, SIGINT, and SIGACTS). In fact the task of TACTICS goes beyond that of an all-source analyst to include the job of a decision-maker, who uses the all-source assessment to select operational courses of action.

6.2 Analytical Systems

Here it is important to acknowledge that raw data (INT reports) are useless for sensemaking, unless some person or system can assign corresponding ***likelihoods*** (discussed above). In the real-world this step is often tacit as an analyst may reason without making his or her estimates of likelihoods explicit. But the fact is that there must be at least an implicit assignment of likelihoods to raw data, if such data are to be of any use in *reasoning to the **most likely** explanation (a hypothesis) or prediction (of evidence).*

The ICArUS challenge problem must make such likelihoods explicit (see Burns, 2014), in order to separate the function of estimating individual likelihoods from the function of aggregating multiple likelihoods. This separation is required, for rigor in measuring various cognitive biases that would otherwise be confounded in experiments.

For example, consider the judgment of Red attack probability, $P_a = P_{t,p,c}$, which is an aggregation of various individual probabilities (P_t, P_p, and P_c) and an input to decision-making. If a human experiment measures only Blue decisions (d or ~d), without measuring the underlying judgments of P_a that affect such decisions, then there is no way to establish if a biased decision stems from

bias in P_a or from biases in other parameters of the decision (e.g., P, U, E_d, $E_{\sim d}$). Likewise, if the contributing judgments (P_t, P_p, P_c) are not measured individually (and collectively), then there is no way to establish if bias in P_a stems from estimating the individual probabilities (and which ones?) and/or aggregating the multiple probabilities (at which stage?).

The same separation is also quite relevant to real-world intelligence because it highlights the computational importance of ***likelihoods***, which are required either implicitly or explicitly to "make sense" of any data. This is especially relevant to the engineering of "systems" that might usefully support sensemaking, as such systems must be able to both compute and communicate likelihoods to human sensemakers (Burns, 2007; 2006).

In fact the distinction between *estimating* individual probabilities and *aggregating* multiple probabilities was the focus of early efforts to design machine systems that could support humans in real-world intelligence and operations functions (see Edwards & Phillips, 1964; Edwards, et al., 1968). Those groundbreaking efforts were aimed at mitigating conservative human biases (Edwards, 1982) by having systems aggregate the likelihoods in tasks of Bayesian inference. Unfortunately the systems were largely unsuccessful in practice, for two reasons.

First, the job of estimating individual likelihoods (needed for input to the aggregation algorithm) was left to human beings, so inputs to the system were subject to human biases of likelihood estimation. Second, and more importantly, it was unrealistic to expect that human beings could and would provide the proper conditional likelihoods needed as input to the system – especially when they did not intuitively understand the aggregation algorithm (Burns, 2007; 2006).

In short, the problem to be solved is not *separation* of the two functions (i.e., *estimation* versus *aggregation*). Rather the problem is *integration* of the two functions – which hinges on communication and coordination whenever the two functions are performed by two different agents (human and system, or human and human, or system and system).

More recently, a prototype system was developed to support humans in performing the integrated functions of likelihood estimation and aggregation. This system, called *Bayesian Boxes* (Burns 2007; 2006), is an interactive visualization using geometric representations of probabilistic information. The system helps humans understand what likelihoods must be estimated, and how they are then aggregated – by intuitively illustrating *what* are the inputs and outputs, as well as *how* the outputs are computed from the inputs. As such, the system is an example of "visual analytics" (see National Research Council, 2013, discussed in Section 6.4.1), which might be implemented, evaluated, and demonstrated in TACTICS.

TACTICS is a useful test-bed in this regard, because it naturally poses the dual problems of estimating likelihoods (from BLUEBOOK knowledge and/or experience) and aggregating those likelihoods with various INT likelihoods from OSINT, HUMINT, and SIGINT. Each of these INT reports is accompanied by an associated probability that quantifies vulnerability, capability, or activity, respectively, akin to the likelihoods that might be developed implicitly or explicitly in real-world *suitability* analyses.

As mentioned in footnotes throughout *Description* (Section 2), each form of suitability analysis might be made more realistic in more complex versions of the basic task. That flexibility makes TACTICS scalable to any level of complexity, ranging from the current "lab" version (which could be further simplified, if desired) to almost any "real" demonstration that might be deemed useful in transition. More realistic demonstrations might be used to portray the integrated challenges of estimating individual likelihoods and aggregating multiple likelihoods, as discussed above.

The lab version of TACTICS developed for Phase 2 is purposely limited with respect to the details of various suitability analyses, for both practical and programmatic reasons. From a practical perspective, if humans were required to perform more detailed suitability analyses themselves, it would detract from the current focus of experiments on sensemaking itself. From a programmatic perspective, more complex and realistic suitability analyses would require human visual perception and natural language processing capabilities, as well as extensive domain expertise (i.e., rich and sophisticated knowledge representations, RASKR), which are all outside the scope of the ICArUS BAA.

6.3 Adversarial "Agents"

Despite limitations noted above, the lab version of TACTICS may hold potential for real-world applications of ICArUS models and insights. This promise stems from adversarial aspects of the task, which serve to make TACTICS:

> *A game of repeated risk assessment and action (Kaplan & Garrick, 1980; Garrick, et al., 2004), posing cognitive challenges that are prototypical of intelligence and operations in threat situations (Burns, 2010; McDonald, 1950) – including counterinsurgency (COIN) and other security domains (airport/border, cyber/network, crime/fraud, drugs/gangs, etc.).*

In particular, a model that plays TACTICS (Blue or Red or both) with human-like biases, as measured and modeled in the lab, may be a useful "agent" in agent-based simulations (Axelrod, 1984; Axelrod, 1997; National Research Council, 1998). Computational simulations are currently performed in many real-world security domains, but the agent models are typically not grounded in psychological or neuro-biological research on cognitive biases. This creates an opportunity for models that are more firmly based on behavioral research, particularly models that can credibly extrapolate from constrained lab conditions (in which they were developed and validated) to real-world situations of interest to the Intelligence Community.

It remains to be seen how well neural models developed by ICArUS can extrapolate to more complex sensemaking (especially given scope limitations of the program, discussed above). Nevertheless, applications may be possible for game situations that involve relatively simple background knowledge and payoff structures, such as the "Stackelberg" game simulations currently being performed to support airport security operations. In fact a recent study in this domain by Pita, et al. (2010) highlights the importance of agent models that can act with human bias, noting that:

"Our results show that the anchoring bias may play an important role in human responses… and exploiting this bias can lead to significant performance improvements. This is an important conclusion… [with] real deployment at LAX and Federal Air Marshals service."

As currently designed in TACTICS, a Blue human plays against a Red agent with a very simple payoff structure for both players. Possible extensions that may prove useful in transition include a Blue agent playing against a Red human, or a Blue agent playing against a Red agent (as in most agent-based simulations, which have no humans in the loop). Extensions might also introduce more complex payoff structures, and/or scale-up complexity along any or all of the six (or more?) types of geospatial intelligence noted above in Section 6.1 – perhaps using "teams" of Blue (and Red) comprising different individuals, each performing different analytical and operational functions but acting together in a coordinated fashion (Powers, et al., 2010).

6.4 Organizational Training

As discussed above, *Adversarial Agents* and *Analytical Systems* are two areas for transition of ICArUS models and insights. A third area that holds potential for transition is *Organizational Training*, based on lessons learned from the design of TACTICS and human/model experiments with the game. Some topics that might be addressed in such a training program are outlined in the following sections.

6.4.1 What is Sensemaking, Anyway?

As a practical matter, the *computational* design of TACTICS (also see Burns, 2014) serves to expose and explain sensemaking more formally than previous research on the topic (see *Introduction*, Section 1). In a first step toward transition, the computational approach has enabled a relational mapping of TACTICS to 26 cases of real-world intelligence, discussed in Section 6.1. This mapping may allow intelligence analysts as well as ICArUS itself to better understand analytic "tradecraft" from the scientific perspective of cognitive computing.

Further steps in the same direction may be informed by knowledge gained in the challenge problem design process, particularly insights associated with cognitive biases (Section 5) and the normative solutions (Section 4) that are required for rigorously measuring and modeling such biases. These insights might be elucidated by a training program that demonstrates biases in hands-on fashion using the current version or tailored demo of TACTICS as a use case.

Perhaps the most important and underappreciated insight of all, which would be made clear in such a demo, is the key role played by ***likelihoods*** – i.e., likelihoods of *evidence* given *hypotheses*, and likelihoods of *hypotheses* given *evidence*. These likelihoods are the critical components of *frames*, or *scripts*, or whatever else one chooses to call the knowledge structures involved in sensemaking.

As discussed in Section 6.2, data are useless for sensemaking without some person or system that infers or assigns associated ***likelihoods***. So tools and techniques for "storing" (warehousing) or "seeing" (visualizing) or "sharing" (disseminating) data are useful for sensemaking only to the extent that they represent *likelihoods* (which most current systems do not) and/or support human

users in estimating and aggregating *likelihoods* (which most current systems do not). This suggests opportunities to advance the practice of intelligence sensemaking (Burns, 2011), where a focus on ***likelihoods*** may lead to novel systems (Section 6.2) as well as future training for the geospatial intelligence workforce.

For example, a recent report by the National Research Council (NRC, 2013) on *"Future U.S. Workforce for Geospatial Intelligence"* begins by stating that:

> *"We live in a changing world with multiple and evolving threats to national security, including terrorism, asymmetrical warfare, and social unrest. Visually depicting and* ***assessing these threats*** *[emphasis added] using imagery and other geographically-referenced information is the mission of the National Geospatial-Intelligence Agency (NGA). As the nature of the threat evolves, so do the tools, knowledge, and skills needed to respond."*

The NRC report reviews *existing* disciplines and core competencies of geospatial intelligence, including those associated with Geographic Information Systems, which are primarily concerned with visually *depicting* various aspects of the threats. The report also identifies ***emerging*** disciplines where new competencies are required for ***assessing*** these threats, including ***"human geography"*** (i.e., understanding the activities of individual and organizations), ***"visual analytics"*** (i.e., cognitive reasoning, especially as aided by visual interfaces), and ***"forecasting"*** (i.e., anticipating outcomes or behaviors using statistics and modeling).

Notice that these new and emerging areas are less concerned with *depicting* aspects of threats and more concerned with ***assessing*** the threats themselves – ultimately to support appropriate actions. As such the emerging disciplines of geospatial intelligence are clearly aligned with the practice of ***sensemaking***, which is concerned with *explaining* (understanding) and *predicting* (forecasting) the behavioral activities of actors in geospatial areas of interest. The more established disciplines of geospatial intelligence are geared more toward developing and depicting data, and performing various forms of suitability (vulnerability, opportunity, capability, etc.) analyses, which in turn serve as inputs to ***threat assessment*** – ***in sensemaking***.

The NRC report goes on to observe that academic degrees and agency training in the emerging disciplines of geospatial intelligence are still in their infancy. Therefore new training programs, like new "tools" (systems, see Section 6.2), represent an opportunity for applying Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking. TACTICS may be useful for that purpose as the game itself is a tool that could be used in training on ***"heuristic and biases"*** (discussed below).

6.4.2 Heuristics and Biases

Although there is much talk of "bias" in the Intelligence Community (e.g., George & Bruce, 2008), most of this talk is not grounded in computational theory or experimental testing, let alone a combination of the two. As a result there are many unsupported arguments about if and when humans are biased or not – and why it is important – and what can be done about it. Conversely, the academic literature contains many computational and experimental studies of cognitive

biases, but each is typically limited to one (or a few) isolated bias(es) studied in the context of an artificial lab task that lacks natural richness. A potential contribution of TACTICS is to help bridge this apparent gap with an adversarial game task combining scientific rigor with analytic relevance.

The design of TACTICS shows first-hand how difficult it is to rigorously define relevant biases, because such definitions are sensitive to assumptions that must be made in computing solutions. For example, perhaps the most infamous bias discussed in intelligence circles is *Confirmation Bias*, and yet the normative analyses of TACTICS show that a *confirmation preference* (in seeking evidence) is actually optimal assuming realistic values for sensor parameters – hence it is not really a "bias" per se. This suggests that other so-called biases may also be useful (if not optimal), too – at least in many situations of real-world importance. In fact much research in recent years points to the advantages of heuristics that are naturally employed in cognitive reasoning (Gigerenzer, 2000; Gigerenzer & Selten, 2001; Gigerenzer & Todd, 1999).

It appears that even the most basic distinction between a "heuristic" (psychological process) and "bias" (behavioral result) is not well-appreciated in the Intelligence Community, despite the influential *Psychology of Intelligence Analysis* (Heuer, 1999). Thus a training program may add value simply by clarifying and exemplifying ***"heuristics and biases"*** in a use case of geospatial intelligence demonstrated by TACTICS. More value could be added by addressing other important definitions and distinctions from a computational perspective, much like the design of TACTICS itself began by formalizing vague notions like "frames" and "re-framing" and "set-shifting" in terms of *hypotheses, evidence, and* ***likelihoods***. This would help relate the emerging view of "sensemaking" to an established view of "hypothesis testing", which has been used by some in the Intelligence Community for decades (Zlotnick, 1970; Fisk, 1972; Schweitzer, 1976; Heuer, 1999) to formalize the practice of the same basic process that ICArUS calls sensemaking.

Beyond these definitions and distinctions, a training program might also distinguish the cognitive processes that can lead to various sensemaking biases, i.e., the processes of inferencing, decision-making, and foraging. For instance, one can postulate a *Confirmation Bias* in combining likelihoods, and/or in assigning likelihoods (which would then be used in combining likelihoods), and/or in selecting evidence (which would then be used in assigning likelihoods and combining likelihoods). Typically authors focus on only one of these behaviors without addressing the others in concert. For example, two recent and relevant experiments on *Confirmation Bias* measured the relative "weight" of support assigned to one or more hypotheses (Lehner, et al., 2008; Lehner, et al., 2009). However, the "overweighting" observed in these studies might be mitigated or even reversed by the "conservative" biases (Edwards, 1982) known to affect human aggregation of such "weights" (which are actually ***likelihoods***).

In some cases it appears that the so-called *Confirmation Bias* may actually be an author's own bias – as the term has come to be a catch-all for almost any favored effect that the author himself would like to "confirm" in lab testing or storytelling. This was illustrated in a formal analysis (Burns, 2005) of a well-known story dealing with so-called *Confirmation Bias* (Perrow, 1984). TACTICS enables more integrated and empirical measures of the "confirming" and "disconfirming" cognitive processes, so that associated biases (whatever they are called) can be studied in a more rigorous and relevant fashion.

Finally, it should be acknowledged that a completely different type of *Confirmation Bias* may in fact be the most ubiquitous and important – and yet it is apparently the least studied of all. That type of *Confirmation Bias* applies to ***creating*** a frame of discernment (set of hypotheses) in the first place, whereby an analyst may tend to confirm one or more of his current hypotheses rather than generate new hypotheses that may better explain the evidence. This is the familiar bias we see in major intelligence failures sometimes referred to as *"failures of imagination"* (The 9/11 Commission Report, 2004).

Unfortunately this bias is difficult to study with computational and experimental rigor. Instead it is easier to study how people reason over a controlled (fixed and known) sets of hypotheses. So, here again, the literature on biases itself may be "biased" in "confirming" what is most convenient to study rather than addressing what is most relevant and important. Admittedly the design of TACTICS also suffers from this same bias, driven by the need to meet BAA requirements for evaluating models in Comparative Performance Assessment (CPA) and Cognitive Fidelity Assessment (CFA). However, Missions 4 and 5 of TACTICS do approach a more creative sensemaking in which humans are making forensic inferences at a higher level of abstraction (e.g., the Red style, Passive or Aggressive) in order to support prognostic inferences at a lower level of abstraction (e.g., the probability of Red attack).

Looking beyond Phase 2, TACTICS might be extended to support future research on more "wicked" (open) problems posing challenges of creative (abductive) sensemaking. These are clearly the problems of most relevance to the Intelligence Community, often explored in "team" training and Red-Blue exercises. But thus far they have not been researched with much rigor using computational models and experimental testing (Powers, et al., 2010; Ambrose & Ahern, 2008). These problems that require "creative thinking" have also not been addressed by research on "critical thinking", which is typically measured by closed-form questions in multiple-choice format (MITRE, 2014).

6.4.3 Structured Analytic Techniques

A final topic that deserves mention, in the context of *Organizational Training*, is Structured Analytic Techniques (SATs). The SATs include Analysis of Competing Hypotheses (ACH, see Heuer, 1999), designed to help address *Confirmation Bias*, as well as many other techniques (Beebe & Pherson, 2012). These SATs are promoted as tools that can mitigate biases and prevent intelligence failures, and they may indeed do so. But it is not clear to what extent SATs actually help, or in what respects SATs may not help and may even hurt.

For example, one empirical study of *Confirmation Bias* (Lehner, et al., 2008) showed that ACH offered a significant reduction in bias only for participants without intelligence analysis experience. Also, results of the ICArUS experiments (using experienced and inexperienced participants) shows that numerous biases remain even when structured techniques like ACH are employed. In Phases 1 and 2, the experimental protocol effectively forced all participants to adopt the technique of ACH, and yet significant biases were still measured in individual and average human responses. This suggests that ACH does not eliminate biases, and it may even introduce biases.

The point here is not to argue for or against the use of ACH or any other SAT. Rather, the point is that much work remains to establish the advantages and disadvantages of SATs, using rigorous and relevant evaluations. Moreover, it should be noted that most "structured" techniques are merely "questions" or "checklists", so they are basically what most analysts (at least experienced analysts) would be doing anyway – implicitly and naturally. This may help explain the limited benefit of ACH noted above (Lehner, et al., 2008), which was found only for participants without intelligence experience.

Moving beyond the questions and checklists of SATs, there appear to be opportunities for more revolutionary advances in analytic tools, techniques, and training. But these advances will require a cognitive-scientific approach that addresses intelligence analysis from a computational perspective (Burns, 2014), like the approach adopted by ICArUS and its challenge problem of TACTICS.

6.5 Conclusion

As noted in the *Introduction* (Section 1), a computational approach is needed to advance the scientific understanding of sensemaking at *functional, psychological, and biological* levels of abstraction. Research products of ICArUS span all three levels, to promote transition in the form of *Analytical Systems, Organizational Training*, and *Adversarial Agents*, as follows:

> At the *functional* level, ***formal design*** of a challenge problem exposes the computational functions of sensemaking, including inferencing (prognostic and forensic), decision-making, and foraging. In that regard, ICArUS holds potential for transition to *Analytical Systems*.
>
> At the *psychological* level, ***human data*** and Bayesian benchmarks enable a deeper understanding of heuristics and biases in geospatial sensemaking. In that regard, ICArUS holds potential for transition to *Organizational Training*.
>
> At the *biological* level, ***neural models*** that emulate human behavior can help explain the fundamental mechanisms that give rise to sensemaking biases. In that regard, ICArUS holds potential for transition to *Adversarial Agents*.

7 Definitions

Abducting is a form of *sensemaking* in which *re-framing* creates new *hypotheses* not previously considered in one's *frame of discernment*.

Bayesian refers to the use of Bayes Rule for updating beliefs in *hypotheses* given *evidence*. Bayes Rule is mathematical specification of how *prior* (before *evidence*) *probabilities* of *hypotheses* and conditional *likelihoods* of *evidence* (given *hypotheses*) are combined to compute *posterior* (after *evidence*) *probabilities* of *hypotheses*. *Bayesian* also refers to the optimal computation of expected utility, in decision-making situations, as the product of probability and utility summed across all possible outcomes of an option.

Causal Hierarchy is an ordering of causal factors in which higher factor(s) cause or constrain lower factor(s), such that: the assumption of a higher factor (*hypothesis*) can be used to infer the probability of a lower factor (*evidence*) – in a *prediction* of *evidence* (i.e., in forward inference); and the observation of a lower factor (*evidence*) can be used to infer the probability of a higher factor (*hypothesis*) – in an *explanation* of *evidence* (i.e., in backward inference). In TACTICS, the causal hierarchy is represented by four arrows (→) as follows*: intent → tactic → action → feature → datum*.

Confidence is a measure of belief in the truth of a *hypothesis* (i.e., *confidence* in *explanation*) or *evidence* (i.e., *confidence* in *prediction*), quantified as a *likelihood* (*probability*) ranging from zero to one. [In a more specific sense, not used here, *confidence* is a measure of second-order *probability*, i.e., the *probability* that some *probability* is correct.]

Evidence is a report of a datum or feature or action or tactic or anything else that might be observed at any level of a *causal hierarchy*. The term *evidence* may be used in referring to actual observations (i.e., *evidence* that may be *explained* by *hypotheses* and *likelihoods*) or potential observations (i.e., *evidence* that may be *predicted* by *hypotheses* and *likelihoods*).

Explanations are backward inferences about the *likelihoods* of *hypotheses* in light of *evidence*.

Frames are knowledge structures, comprising *hypotheses, evidence,* and *confidences*, including conditional *likelihoods* of *evidence* (i.e., conditional on *hypotheses*) as well as conditional *likelihoods* of *hypotheses* (i.e., conditional on *evidence*). In *spatial context frames, likelihoods* depend on spatial factors. In *event sequence frames, likelihoods* depend on temporal (and spatial) factors.

Frame of Discernment refers to the set of *hypotheses* (and/or set of *evidence*) over which one reasons and assigns *confidence*.

Hypotheses are possible *explanations* of *evidence*, typically involving causal reasons for *evidence*.

Inferencing is the assignment of *confidences* to *hypotheses* in one's *frame of discernment*. *Abducting* is a class of *inferencing* that involves the creation of new *hypotheses*.

Likelihood is a general term referring to *confidence* measured by *probability*. The term *likelihood* is also used in a more specific (*Bayesian*) sense when referring to the *probability* of some *evidence* conditional on a *hypothesis*.

Posterior refers to the result of *Bayesian* updating, in which *prior probabilities* are updated with *likelihoods* (of *evidence* given *hypotheses*) to compute *posterior probabilities* (of *hypotheses* given *evidence*).

Predictions are forward inferences about the *likelihoods* of *evidence* in light of *hypotheses*.

Prior refers to the *probability* of a *hypothesis* in the absence of *evidence*, i.e., *prior* to obtaining the *evidence*.

Probability is a mathematical measure of belief in the truth of a *hypothesis* or *evidence*. As such, *probability* is a measure of mental *confidence*.

Re-framing **(aka** ***Set-shifting*****)** is a revision of *hypotheses*, or revision of *confidences* across *hypotheses*, in which the most likely *hypothesis* changes due to the observation of surprising *evidence* (i.e., *evidence* that is not likely to be caused by the currently-most-likely *hypothesis* or *hypotheses*).

Sensemaking is a recurring cycle of obtaining *evidence* and updating *confidence* in competing *hypotheses*, to *explain* and *predict* an evolving situation.

Set-shifting is another term for *re-framing*.

Spatial Hierarchy is an ordering of spatial features in which higher level(s) include features at lower level(s). In TACTICS, an area of interest includes regions, and a region includes circles around points – thus the spatial hierarchy is: area(region(circle(point))).

Temporal hierarchy is an ordering of temporal events in which higher level(s) include events at lower level(s). In TACTICS, a mission is a sequence of batches, and a batch is a sequence of trials. Each trial includes a sequence of temporal-spatial features (of events, from INT reports), in stages of the trial, thus the temporal hierarchy is mission(batch(trial(stage))).

8 References

Ambrose, F., & Ahern, B. (2008). Unconventional red teaming. In Chesser, N. (ed.), *Anticipating Rare Events: Can Acts of Terror, Use of Weapons of Mass Destruction, or Other High Profile Acts be Anticipated?,* pp, 136-139, http://redteamjournal.com/papers/U_White_Paper-Anticipating_Rare_Events_Nov2008rev.pdf

Axelrod, R. (1997). *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton, NJ: Princeton University Press.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

BAA (2010). IARPA Broad Agency Announcement, *Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking (ICArUS*). IARPA-BAA-10-04, April 1, 2010.

Barlett, F. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge, UK: Cambridge University Press.

Beebe, S., & Pherson, R. (2012). *Case Studies in Intelligence Analysis: Structured Analytic Techniques in Action*. Los Angeles, CA: Sage.

Berbaum, K., Franken, E., Dorfman, D., Rooholamini, S., Kathol, M., Barloon, T., Behlke, F., Sato, Y., Lu, C., & el-Khoury, G. (1990). Satisfaction of search in diagnostic radiology. *Investigative Radiology, 25(2),* 133-140.

Berg, E. (1948). A simple objective technique for measuring flexibility in thinking. *Journal of General Psychology, 39,* 15-22.

Burns, K. (in press). Computing the creativeness of amusing advertisements: A Bayesian model of Burma-Shave's muse. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*.

Burns, K. (2014). *A Computational Basis for ICArUS Challenge Problem Design.* MITRE Technical Report, MTR140415.

Burns, K. (2012). EVE's energy in aesthetic experience: A Bayesian basis for haiku humor. *Journal of Mathematics and the Arts, 6,* 77-87.

Burns, K. (2011). The challenge of iSPIED: intelligence sensemaking to prognosticate IEDs. *The International C2 Journal, 5(1),* 1-36.

Burns, K. (2010). Strategic style in pared-down poker: With applications to terror networks and systems failures. In Argamon, S., Burns, K., & Dubnov, S. (eds.), *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning*. Berlin: Springer.

Burns, K. (2007). Dealing with probabilities: On improving inferences with Bayesian Boxes. In Hoffman, R. (ed.), *Expertise Out of Context*. New York: Lawrence Erlbaum, pp. 263-280.

Burns, K. (2006). Bayesian inference in disputed authorship: A case study of cognitive errors and a new system for decision support. *Information Sciences, 176*, 1570-1589.

Burns, K. (2005). Mental models and normal errors. In Montgomery, H., Lipshitz, & Brehmer, B. (eds.), *How Professionals Make Decisions*. Mahwah, New Jersey: Lawrence Erlbaum.

Burns, K., & Bonaceto, C. (2014). *Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking (ICArUS): Phase 2 Challenge Problem Walkthrough.* MITRE Technical Report, MTR140414.

Burns, K., Greenwald, H., & Fine, M. (2014). *ICArUS Phase 1 Challenge Problem Design and Test Specification*. MITRE Technical Report, MTR140410.

Burns, K., Fine, M., Bonaceto, C., & Oertel, C. (2014). *Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking (ICArUS): Overview of Test and Evaluation Materials*. MITRE Technical Report, MTR140409.

Burns, K., & Demaree, H. (2009). A chance to learn: On matching probabilities to optimize utilities. *Information Sciences, 179*, 1599-1607.

Davis, M. (1997). Game *Theory: A Nontechnical Introduction*. New York: Dover.

Edwards, W. (1982). Conservatism in human information processing. In Kahneman, D., Slovic, P., & Tversky, A., (eds.), *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press, pp. 359-369.

Edwards, W. (1961). Behavioral decision theory. *Annual Review of Psychology, 12*, 473-498.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin, 51(4)*, 380-417.

Edwards, W., Phillips, L., Hayes, W., & Goodman, B. (1968). Probabilistic information processing systems: Design and evaluation. *IEEE Transactions on Systems, Man, and Cybernetics, 4(3)*, 248-265.

Edwards, W., & Phillips, L. (1964). Man as transducer for probabilities in Bayesian command and control systems. In Shelly, M., & Bryan, G. (eds.), *Human Judgments and Optimality*. New York: Wiley.

Edwards, W., Lindman, H., & Savage, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70(3)*, 193-242.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90(3)*, 239-260.

Fisk, C. (1972). The Sino-Soviet border dispute: A comparison of the conventional and Bayesian methods for intelligence warning. *Studies in Intelligence, 16(2),* 53-62.

Garrick, B., Hall, J., Kilger, M., McDonald, J., O'Toole, T., Probst, P., Parker, E., Rosenthal, R., Trivelpiece, A., Van Arsdale, L., & Zebroski, E. (2004). Confronting the risks of terrorism: Making the right decisions. *Reliability Engineering and System Safety, 86(2),* 129-176.

George, R., & Bruce, J. (2008). *Analyzing Intelligence: Origins, Obstacles, and Innovations*. Washington, DC: Georgetown University Press.

Gigerenzer, G. (2000). *Adaptive Thinking: Rationality in the Real World*. Oxford, UK: Oxford University Press.

Gigerenzer, G. & Selten, R. (2001). *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.

Gigerenzer, G., & Todd, P. (1999). *Simple Heuristics that Make Us Smart*. Oxford, UK: Oxford University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge, UK: Cambridge University Press.

Grabo, C. (2004). *Anticipating Surprise: Analysis for Strategic Warning*. Lanham, MD: University Press of America.

Heuer, R. (1999). *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, CIA.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.

Kaplan, S., & Garrick, B. (1980). On the quantitative definition of risk. *Risk Analysis, 1(1),* 11-27.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94(2),* 211-228.

Klein, G., Phillips, J., Rall, E., & Peluso, D. (2007). A data-frame theory of sensemaking. In Hoffman, R. (ed.), *Expertise Out of Context*. New York: Lawrence Erlbaum, pp. 113-155.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22,* 79-86.

Lehner, P., Adelman, L., DiStasio, R., Erie, M., Mittel, J., & Olson, S. (2009). Confirmation bias in the analysis of remote sensing data. *IEEE Transactions on Systems, Man, & Cybernetics – Part A: Systems and Humans, 39(1),* 218-226.

Lehner, P., Adelman, L., Cheikes, B., & Brown, M. (2008). Confirmation bias in complex analysis. *IEEE Transactions on Systems, Man, & Cybernetics – Part A: Systems and Humans, 38 (3),* 584-592.

Louis, M. (1980). Surprise and sensemaking: What newcomers experience in entering unfamiliar organizational settings. *Administrative Science Quarterly, 25(2),* 226-251.

Macknik, S., King, M., Randi, J., Robbins, A., Teller, Thompson, J., & Martinez-Conde, S. (2008). Attention and awareness in stage magic: Turning tricks into research. *Nature Reviews Neuroscience, 9,* 871-879.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* New York: Freeman.

McDonald, J. (1950). *Strategy in Poker, Business, & War.* New York: Norton.

MITRE (2014). *Critical Analytical Thinking Skills Pilot Test*.

MITRE (2013). *Geospatial Intelligence: A Cognitive Task Analysis. Part 2: Descriptive Task Analysis.*

Monchi, O., Petrides, M., Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin card sorting revisited: Distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *The Journal of Neuroscience, 21(19),* 7733-7741.

National Research Council (2013). *Future U. S. Workforce for Geospatial Intelligence*. Washington, DC: The National Academies Press.

National Research Council (1998). *Modeling Human and Organizational Behavior: Application to Military Simulations*. Washington, DC: National Academy Press.

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2(2),* 175-200.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge, UK: Cambridge University Press.

Perrow, C. (1984). *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books.

Pita, J., Jain, M., Tambe, M., Ordonez, F., & Kraus, S. (2010). Robust solutions to Stackelberg games: Addressing bounded rationality and limited observations in human cognition, *Artificial Intelligence, 174,* 142-1171.

Powers, E., Stech, F., & Burns, K. (2010). A behavioral model of team sensemaking. *The International C2 Journal, 4(1),* 1-10.

Roberts, N., Vesely, W., Haasl, D., & Goldberg, F. (1981). *Fault Tree Handbook*. NUREG-0492, U.S. Nuclear Regulatory Commission.

Schank, R., & Abelson, R. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. Mahwah, NJ: Lawrence Erlbaum.

Schweitzer, N. (1976). Bayesian analysis for intelligence: Some focus on the Middle East. *Studies in Intelligence, 20(2),* 31-44.

Shannon, C., & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.

Steinberg, A., & Bowman, C. (2004). Rethinking the JDL fusion layers. *Data Fusion and Resource Management Architecture at the AIAA Intelligent Systems Conference.*

The 9/11 Commission Report (2004). *Final Report of the National Commission on Terrorist Attacks Upon the United States.*

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.

von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Weick, K. (1995). *Sensemaking in Organizations*. Thousand Oaks, CA: Sage.

Zlotnick, J. (1970). Bayes' theorem for intelligence analysis. *Paper presented at the Conference on the Diagnostic Process,* June 18, Ann Arbor, MI.