MTR140410

MITRE TECHNICAL REPORT

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

Phase 1 Challenge Problem Design and Test Specification

Kevin Burns
Hal Greenwald
Michael Fine

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 14-2189

McLean, VA

Abstract

Phase 1 of the IARPA program ICArUS (Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking) requires a challenge problem that poses cognitive challenges of geospatial sensemaking (BAA, 2010). The problem serves as a modeling challenge for performers and enables assessment in T&E (Test and Evaluation) per BAA guidelines. This document describes the Phase 1 challenge problem and outlines the T&E approach for evaluating models in Neural Fidelity Assessment, Cognitive Fidelity Assessment, and Comparative Performance Assessment.

Note: This document was originally prepared and delivered to IARPA in December, 2011, to support ICArUS Phase 1 T&E efforts that concluded in December, 2012.

Table of Contents

List of Tables

1 Introduction

This document was originally prepared and delivered to IARPA in December, 2011, to support ICArUS Phase 1 Test & Evaluation (T&E) efforts that concluded in December, 2012. Further background is provided in a summary document (Burns, et al., 2014) titled *ICArUS: Overview of Test and Evaluation Materials*, available at http://www.mitre.org/publications.

The ICArUS Phase 1 challenge problem, dubbed AHA (*Abducting Hot-spots of Activity*), poses cognitive challenges prototypical of geospatial sensemaking. The design is informed by reviews of GEOINT doctrine (MITRE, 2012) and interviews with intelligence analysts in various agencies (MITRE, 2013). GEOINT analysis exploits geospatial data and requires inferential reasoning to make sense of an uncertain situation. This includes explaining and predicting identities and activities, based on knowledge and data about causes and effects. *Sensemaking* is a process of inferring the relative likelihoods of hypotheses, given evidence, where the hypotheses may refer to causes and/or effects (MITRE, 2011).

In AHA, the causes are enemy groups that attack in the context of counterinsurgency (COIN) operations. The effects are attacks at sites in an area of interest. The problem is to infer the relative likelihoods of hypotheses about the groups or sites, given evidence in one of two different conditions. In the first condition, evidence is accumulated over time from reports of individual attacks. These reports are deterministic because the groups and sites of past attacks are assumed to be known with certainty. In the second condition, evidence about a single attack is provided by multiple independent sources of data presented one at a time. These data are accompanied by probabilistic knowledge (in the form of rules) needed to infer the relative likelihoods of hypotheses. The two conditions are captured in six tasks[1], see Table 1. Tasks 1-3, which reflect the first condition, require statistical learning. Tasks 4-6, which reflect the second condition, require rule-based sensemaking.

Consistent with Table 3 of the ICArUS Broad Agency Announcement (BAA, 2010), all tasks of the Phase 1 challenge problem involve *spatial context frames* with underlying probabilities that are constant in time. A spatial context frame (MITRE, 2011) includes a hypothesis H (e.g., enemy group) and conditional likelihoods P(e|H) of evidence given the hypothesis, for various geospatial features of evidence e_1, e_2, e_3, etc. For Tasks 1-3 the subject must learn P(H) and P(e|H) from a cumulative statistical sample, in order to make inferences of P(H|e). For Tasks 4-6 the evidence is from multiple independent sources, and the likelihoods P(e|H) needed for sensemaking are provided to the subject via probabilistic rules. These rules represent the causal knowledge (i.e., spatial context frames) required to infer P(H|e) from a body of evidence e = {e_1, e_2, e_3,...} across a set of hypotheses H = {H_1, H_2, H_3,...}.

[1] An additional task, Task 7, is intended to serve as a bridge between Phases 1 and 2 as well as to provide a general platform for constructing more complex, naturalistic sensemaking tasks in future phases. Although human behavioral data will be collected for all seven tasks in Phase 1, Task 7 will not be used for Phase 1 evaluations. The remainder of this document pertains to Tasks 1-6. The Task 7 description is included as Appendix A.

Table 1: Summary of six tasks in the Phase 1 challenge problem.

Task*	Data	Inference	Decision	Feedback	Judgment
1	SIGACTS	% for {A, B}	Pick A or B	A or B	Surprise
2	SIGACTS	% for {A, B, C, D}	Draw four circles Pick A, B, C, or D	A, B, C, or D	Surprise
3	SIGACTS	% for {A, B, C, D}	Click four centers Pick A, B, C, or D	A, B, C, or D	Surprise
4	SIGACT HUMINT SOCINT	% at {1, 2, 3, 4) % at {1, 2, 3, 4)	% at 1, 2, 3, and 4	1, 2, 3, or 4	Surprise
5	SIGACT HUMINT IMINT MOVINT SIGINT SOCINT	% for {A, B, C, D} % for {A, B, C, D} % for {A, B, C, D} % for {A, B, C, D}	% for A, B, C, and D	A, B, C, or D	Surprise
6	SIGACT HUMINT 3 of 4: IMINT MOVINT SIGINT SOCINT	% for {A, B, C, D} % for {A, B, C, D} % for {A, B, C, D} % for {A, B, C, D}	Choose next INT % for A, B, C, and D	A, B, C, or D	Surprise

***Tasks 1-3 involve statistical learning; Tasks 4-6 involve rule-based sensemaking.**

For all six tasks, likelihoods are conditioned only on spatial features of evidence in a scene – i.e., spatial context frames (for Phase 1, per BAA Table 3). None of the tasks involve temporal dependencies in which likelihoods are conditioned on temporal features of evidence in a sequence – i.e., *event sequence frames* (for Phase 2, per BAA Table 3). In Tasks 1-3, likelihoods are learned over time but the generative probabilities are constant in time, and all trials are independent (i.e., no temporal dependencies). In Tasks 4-6, evidence is presented in stages of a trial, but all evidence on a given trial is independent of time within the trial, and all trials are independent (i.e., no temporal dependencies).

2 Description

Consistent with the cognitive challenges of GEOINT sensemaking, each task of the challenge problem requires inference under uncertainty given geospatial data from one or more sources. The six tasks are described in Table 2.

Table 2: Description of six tasks in the Phase 1 challenge problem.

Task	Description
1	**Learning two groups**. [10 blocks of 10 trials]. Attacks by two groups (A, B) are generated (SIGACTS), one on each trial, by two normal distributions – each with a different base rate, center location, and dispersion. On the last trial of each block the subject is shown an attack location and asked to predict the probability of attack by each group, {A%, B%}. The subject then makes a forced choice (A or B), to allocate resources against the attack. At the end of the trial the subject sees ground truth (A or B) at the attack location and he reports surprise (on a 7-point Likert scale).
2	**Learning four groups**. [5 blocks of 20 trials]. Unlike Task 1, there are four different groups (A, B, C, and D). Also, before reporting {A%, B%, C%, D%}, the subject is required to draw four circles – each representing the “two-to-one boundary” for a group (i.e., attack by the group is twice as likely inside the circle compared to outside the circle). After reporting the probabilities {A%, B%, C%, D%} and making a forced choice in allocating resources against one group, the subject is shown ground truth (A, B, C, or D) at the attack location and he reports surprise.
3	**Finding centers**. [5 blocks of 20 trials]. Unlike Task 2, the task is to find group centers along roads “as the cow walks” (rather than "as the crow flies"). After each block of trials, the subject clicks four points A, B, C, and D on the roads – each representing the “center of gravity” for a group (i.e., maximum likelihood location for a group’s center, given the group’s attacks). The subject then reports {A%, B%, C%, D%} at the attack location. After reporting the probabilities and making a forced choice in allocating resources, the subject is shown ground truth (A, B, C, or D) at the attack location and he reports surprise.
4	**Scoping attacks**. [10 trials, 2 stages per trial]. Unlike Task 3, the subject is given the center (from HUMINT) for one group, along with four possible locations 1, 2, 3, and 4 of attack by that group. In stage 1, the subject is required to estimate the probability of attack at each location based on a normal distance decay function along roads. In stage 2, the subject is given the regional boundaries for groups (SOCINT), along with the inside/outside region attack likelihoods from PROBS. The subject is required to aggregate the prior (HUMINT) and SOCINT probabilities. The subject allocates resources in percentages (not a forced choice) across the sites, and is shown ground truth (1, 2, 3, or 4). The subject reports surprise after observing ground truth.
5	**Fusing layers**. [10 trials, 4 stages per trial]. Unlike Task 4, the subject is given an attack location (SIGACT) along with group centers (HUMINT) for four groups (A, B, C, D). The subject is also given the probabilities of attack by each group {A%, B%, C%, D%}, based on the HUMINT distance function. The subject is then given four more INTS (IMINT, MOVINT, SIGINT, SOCINT), one at a time, along with the likelihoods (from PROBS) for features of INTS. At each stage the subject updates probabilities {A%, B%, C%, D%}. After the last stage the subject allocates resources in percentages {A%, B%, C%, D%} against the groups. At the end of the trial the subject is shown ground truth and he reports surprise.
6	**Choosing layers**. [10 trials, 3 stages per trial]. Much like Task 5, except the subject decides which layer of INT (IMINT, MOVINT, SIGINT, or SOCINT) to receive on each of three stages, after receiving the initial HUMINT. The criterion for each choice is to maximize the expected amount of useful information to be gained from the choice of an INT.

In all tasks, stimuli and responses are designed to minimize the role of language processing, visual perception, and general background knowledge known as RASKR (rich and sophisticated knowledge representations) – in accordance with BAA Appendix E.

For statistical learning in Tasks 1-3, data are provided in the form of SIGACTS (significant activities) reporting the enemy group and site of each attack. The problem is to infer, for a given attack location, the probability that each group (i.e., each H) is responsible for the attack at that location – based on evidence accumulated over past attacks by the various groups at various sites. Tasks 1-3 avoid issues associated with human RASKRs by giving subjects (humans and models) the data they need in a cumulative statistical sample of past attacks.

For rule-based sensemaking in Tasks 4-6, diverse sources (INTS) provide one datum at a time in stages of a trial. The problem is to update beliefs at each stage of the trial, based on probabilistic rules (PROBS). The data include HUMINT (human intelligence), IMINT (image intelligence), MOVINT (movement intelligence), SIGINT (signal intelligence), and SOCINT (socio-cultural, aka *human terrain*, intelligence). Tasks 4-6 avoid issues associated with human RASKRs by giving subjects the input they need in the form of probabilistic rules (PROBS) and deterministic data (INTS).

All tasks provide deterministic feedback at the end of each trial, and this feedback serves to score a subject's performance on the trial (see *Feedback* below). In Tasks 1-3 the feedback also serves as data for statistical learning across trials of a task, thereby providing tight coupling between training and testing. For Tasks 4-6, sensemaking is accomplished via probabilistic rules that are given to subjects and are held constant across tasks and trials, as contrasted to statistical learning from deterministic feedback in Tasks 1-3.

Table 2 provides short descriptions of Tasks 1-6. The descriptions use terms like "groups" and "sites" and "attacks" and "centers" in order to maintain generality, as these terms and the basic concepts that underlie them apply to many domains of intelligence analysis.

More concretely, the design of AHA was informed by the challenges of counterinsurgency (COIN) intelligence and operations. In that case the groups are insurgents with differing resources and preferences that lead to different patterns of attack, where a pattern includes various spatial features that are likely or unlikely for each group. A group "center" might correspond to a safe house, a weapons cache, or some other type of physical resource or operations center. With respect to a subject's operational decision to allocate resources, the resources might correspond to troops that are assigned against groups or sites in offensive or defensive operations. However, in human experiments with AHA, the use of "loaded" terminology will be avoided to reduce any effects arising from human RASKRs – especially among subjects who may have knowledge of COIN or other domains.

The responses to be reported by subjects are specified in Table 2, and the rules (PROBS) are summarized in Table 3 (see *Specification*, Section 3 below). As shown in Table 1, each task involves inference, decision, feedback, and judgment – which are all discussed further below.

2.1 Inference

For each task of AHA, the inference is a subjective estimate of probabilities across a set of hypotheses. These probabilities are the primary sensemaking responses to be used in assessing model predictions of human performance (see *Evaluation*, Section 4).

In Tasks 1, 2, 3, 5, and 6, the hypotheses refer to enemy groups (e.g., G = A, B, C, or D) and the evidence includes the location of an attack site (Z), such that the inference of P(H|e) is an inference of P(G|e(Z)). For Task 4, the evidence includes a known group and the hypotheses refer to various sites (e.g., Z = 1, 2, 3, or 4) at which the group might attack, such that the inference of P(H|e) is an inference of P(Z|e(G)).

For each task, a trial of testing begins with deterministic data from a SIGACT report on the group or site of an attack. This datum along with other INTS (in Tasks 4-6) are the evidence from which probabilities of hypotheses are inferred on a trial – using likelihoods that are learned (in Tasks 1-3) or known (in Tasks 4-6). When the site Z is given by SIGACT (in Tasks 1, 2, 3, 5, and 6) then the task is to infer P(G|e(Z)) for all groups. When the group G is given by SIGACT (in Task 4) then the task is to infer P(Z|e(G)) for all sites.

In Tasks 4-6, deterministic data from INTS are useful for sensemaking only in conjunction with probabilistic knowledge about the likelihoods of data features in light of various hypotheses. The challenge problem provides subjects with this knowledge in the form of probabilistic rules (PROBS). If these rules were not given to subjects then the likelihoods would have to be assumed by subjects in order to accomplish sensemaking, and subjects would need to report all the likelihoods that were assumed at each stage of each trial of each task in order for T&E to assess performance. By giving the rules in PROBS, we reduce the burden on subjects and ensure that all humans (and models) use the same likelihoods. This enables T&E to measure average sensemaking performance across human subjects, as required by the BAA.

The rules are designed to reflect causal factors, typical of real-world activities that produce evidential effects observed in geospatial data. Some causal factors apply equally to all groups, such as the PROBS for HUMINT and SOCINT – both of which reflect general constraints of cost and risk that would apply to any group. For example, per PROBS for HUMINT, the probability of attack decreases with distance from the group center, and this is consistent with the type of cost (distance decay) function found in many real-world GEOINT applications (MITRE, 2013). Similarly, per PROBS for SOCINT, the probability of attack by a group is highest within that group's region, and this is consistent with a risk function whereby a group incurs less chance of failure within its region of influence. Other causal factors, more specific to the attack style of each group, are captured in the PROBS for IMINT and MOVINT. For example, the groups vary in their preferences for attacking near government or military buildings, and in their preferences for attacking in dense or sparse traffic.

Although there is no guarantee that subjects will use the causal rules of attack patterns, as given in PROBS, we believe it is reasonable to assume they will do so, for four reasons. First, the rules are numerically simple and graphically displayed in order to facilitate understanding and usability by subjects. Second, all stimuli are purposely limited to only those features of INTS that appear in rules of PROBS, thereby minimizing any effects from RASKRs that humans might bring to the task. Third, without using the rules of PROBS a subject would have no basis for making sense of the data (INTS). Fourth, the statistics of stimuli (INTS) presented across trials of each task will be designed to reflect the structure of the rules (PROBS) at least approximately, such that a subject will have no experiential basis not to trust and use the rules.

Even when all rules are given in PROBS, the sensemaking problem is still far from trivial for human subjects. This is because multiple INTS/PROBS must be aggregated in order to make inferences about the relative likelihoods of various hypotheses.

2.2 Decision

Each task involves one or more types of decision, in order to measure aspects of cognition that are not captured by the primary measure of sensemaking discussed above under *Inference*. These decisions also serve to engage human participants in experiments by giving them a clear purpose for sensemaking.

In Task 1 the decision is a forced choice, i.e., to allocate resources against attacks by one of two groups. Tasks 2-3 involve a similar forced choice, but the choice is among four groups rather than two groups. Tasks 2-3 also require simple actions to measure parameters not made explicit by subjects' reports of probabilities across hypotheses. In Task 2 the action is to draw a circle (for each group), which measures two parameters of a 2-D spatial distribution: a center location and dispersion parameter (diameter). In Task 3 the action is to click the location believed to be the center (for each group) of a 1-D spatial distribution along a road. Tasks 4-6 involve decisions whereby subjects allocate resources in percent values against groups or sites, based on the probability distributions reported in sensemaking. These decisions are designed to test for Probability Matching (see *Evaluation*, Section 4). Task 6 also includes choices of data (INTS) in stages of each trial, and these decisions are designed to test for Confirmation Bias (see *Evaluation*, Section 4).

Although the actions in Tasks 2-3 involve movements (draw, click) for human subjects, these actions are only a means for collecting data on underlying parameters associated with subjective probability distributions. As such the actions do not require motor function (which is out of scope per BAA Section 1.A.6) and performers' models should provide analogous responses.

2.3 Feedback

Deterministic feedback on the group or site of attack, from "ground truth", is provided at the end of each trial of each task. For Tasks 1-3, this feedback (along with data from training trials between testing trials) provides data (SIGACTS) needed for statistical learning of likelihoods.

For Tasks 4-6 there are far too few trials to acquire knowledge of likelihoods via statistical learning. The same is true in real-world intelligence, where most of an analyst's knowledge about enemy patterns of behavior is acquired through reports developed by other analysts. Thus probabilistic rules (PROBS) are provided as input for use across all trials of Tasks 4-6, although deterministic feedback on ground truth is still provided at the end of each trial in Tasks 4-6 as it is in Tasks 1-3.

Feedback on ground truth, provided at the end of each trial, offers one way to score sensemaking and decision-making performance on each task. For example, assume each trial of each task has a maximum score of 100 points, and assume that on a single trial of a task the subject assigns resources in percentages {40%, 30%, 20%, 10%} to groups {A, B, C, D}, respectively. If ground truth is "B" then the score, call it S2, would be 30 points. This score is referred to as a "secondary" score (and denoted S2) because it reflects elements of luck and skill – and because it measures decision-making performance rather than sensemaking performance directly.

A "primary" score (denoted S1) that measures skill in sensemaking will also be computed. This score is a measure of similarity S% (see *Evaluation*, Section 4) between subjective and normative probability distributions, where S% = 100% will produce a score of S1 = 100 points and any other S% will produce a score of S1 = S points. The score S1 will be computed on each trial but will only be provided to a subject as an average over trials of a task, after all trials for the task have been completed.

For example, assume that a subject's probabilities reported in sensemaking are {40%, 30%, 20%, 10%}, and assume that these are equal to the normative probabilities. Also assume that the subject makes the normative decision and assigns 100% of resources to group A. The ground truth will likely not be "A" (i.e., because not-A has probability of 60%), e.g., ground truth of "B" would result in a score of S2 = 0 points. However, the same trial would produce a score of S1 = 100 points. This dual scoring scheme, using S1 and S2, is designed to measure pure skill (S1) in sensemaking as well as the blend of luck and skill (S2) in decision-making. S2 is included because it is representative of the feedback typically received in real-world situations, where decisions are made based on probabilistic inferences (in sensemaking) and yet feedback comes in the form of deterministic data. That is why subjects are given feedback on S2 after every trial. S1 is included because the focus of ICArUS is intelligence sensemaking (not operational decision-making) and S1 is a more rigorous and relevant measure of sensemaking.

With respect to S2, there is a difference regarding how resources are allocated in Tasks 1-3 versus Tasks 4-6. In Tasks 1-3, the subject makes a forced choice among options (groups). If the choice matches ground truth then the subject scores S2 = 100 points on the trial, and if the choice does not match ground truth then the subject scores S2 = 0 points on the trial. In Tasks 4-6, the subject allocates resources in percentages (0-100%) across all options (groups or sites). One option (group or site) then appears as ground truth, and the subject's percent assigned to that option is the S2 point score for the trial.

In sum: S1 is a score of intelligence sensemaking, provided to the subject only after all trials of a task are complete. This feedback on S1 (at the end of a task) is primarily to keep human subjects engaged in the tasks by giving credit for skill without the element of luck. S2 is a score of operational decision-making provided at the end of each trial. This feedback is also intended to keep human subjects engaged, as well as to simulate the conditions under which feedback (which includes an element of luck) is typically provided to analysts and operators – i.e., by an outcome that is observed after a decision has been made. The allocation of resources (used to score S2) is an all-or-nothing forced choice in Tasks 1-3, and a percent across options in Tasks 4-6.

Note that feedback from S1 cannot directly affect a subject's performance on any task because S1 is provided only after all trials of a task are complete. However feedback from S2 on one trial of a task can affect performance on future trial(s) of that task, depending on how the subject deals with the feedback. This is by design, to capture effects of deterministic experiences on probabilistic inferences. To the extent that such effects do occur, they will likely be related to subjective feelings of surprise that arise from the mismatch between what is expected and what is experienced. Therefore the challenge problem will also obtain a subjective measure of surprise from human subjects, as discussed under *Judgment* below.

2.4 Judgment

After receiving ground truth, on each trial of each task, the subject will be asked to report *surprise* on a 7-point Likert scale. Although performers' models are not required to report surprise for any of the assessment components per BAA (see *Evaluation*, Section 4), a model's ability to predict human reports of surprise may add to the credibility of the model.

3 Specification

The Phase 1 challenge problem is further specified here by briefly describing how stimuli, in the form of SIGACTS and various INTS, will be generated for Tasks 1-6.

For Tasks 1-2, the generative model is a symmetric 2-D Gaussian with different center location and dispersion parameter for each group. Each group will also have a different base rate at which the Gaussian is sampled for generating attack locations. Group center locations and dispersion parameters will be chosen to make the problem feasible for subjects and assessable in T&E. Designers' judgments on these matters will be tested in a pilot experiment with results used to refine stimuli for the final experiment of Phase 1. For Task 3, the Gaussian (which is the same for all groups) will be 1-D with distance measured along roads. The road network in Task 3 (as well as in Tasks 4-6) will comprise ~4 roads. Care will be taken in designing the stimuli to eliminate the need for fine-scale perceptual judgments of distance.

For Tasks 4-6, the problem is rule-based sensemaking using data (INTS) and rules (PROBS). Table 3 specifies the values of PROBS for the features of all INTS, including HUMINT, IMINT, MOVINT, SIGINT, and SOCINT. SIGACTS are deterministic reports of the attack site and/or group, so there are no SIGACTS rules in PROBS.

As specified in Table 3, each INT depicts a single feature that is multi-valued, either continuous (e.g., distance in HUMINT) or discrete (e.g., dense traffic vs. sparse traffic in MOVINT). The likelihoods are given numerically by PROBS for the associated INT. Note that all INTS are independent, such that the likelihoods for a given INT are not conditionally dependent on the features of any other INTS. Note also that INT features depict only spatial context such that the PROBS reflect spatial context frames – so there are no temporal dependencies within or between INTS (i.e., no event sequence frames).

Table 3: Probabilistic if-then rules (PROBS) for specific types of data (INTS).

INTS	PROBS
HUMINT	**If** a group attacks, **then** the likelihood is a Gaussian function (σ = 10 miles) of distance along road(s) from the group's center.
IMINT	**If** the attack is near a government building, **then** attack by A or B is four times as likely as attack by C or D. **If** the attack is near a military building, **then** attack by C or D is four times as likely as attack by A or B.
MOVINT	**If** the attack is in dense traffic, **then** attack by A or C is four times as likely as attack by B or D. **If** the attack is in sparse traffic, **then** attack by B or D is four times as likely as attack by A or C.
SIGINT	**If** SIGINT on a group reports chatter, **then** attack by that group is seven times as likely as attack by each other group. **If** SIGINT on a group reports silence, **then** attack by that group is one-third as likely as attack by each other group.
SOCINT	**If** the attack is in a group's region, **then** attack by that group is twice as likely as attack by each other group.

In Table 3, the likelihoods (PROBS) are presented in one of two forms. For HUMINT, likelihoods are continuous and of the form P(e|H), where e is distance along a road (from a group's center) and H is the group center. For IMINT, MOVINT, SIGINT, and SOCINT, the likelihoods are discrete and presented in the normalized form of $P(H|e) = P(e|H) / \Sigma_h P(e|H_h)$, where index h refers to elements in the set of hypothesized groups $\{H_h\} = \{A, B, C, D\}$. These normalized likelihoods are also depicted in bar graphs that are shown to subjects when the feature (e) of each INT is received on each stage of each trial. In essence, the bar graphs depict the probabilities of hypotheses given a datum e, $\{P(H_h|e)\}$, assuming the datum e was the only evidence – and assuming a uniform prior distribution. Given a datum e and associated bar graph, the subject's task is to combine the distribution $\{P(H_h|e)\}$ with his prior distribution $\{P(H_h)\}$, which is also displayed in a bar graph, in order to compute and report a posterior (updated) distribution $\{P(H_h|e)\}$.

Stimuli for Tasks 4-6 will be generated in the form of INT features for each trial. The experimental design will generate a reasonably diverse set of stimuli (i.e., different road networks, group centers, etc.) while at the same time satisfying the rules of PROBS at least approximately. No attempt will be made to match the PROBS exactly, because relatively small samples based on relatively few trials (as we have in Tasks 4-6) would not be expected to match PROBS exactly. However the magnitudes of likelihoods will be reflected at least approximately, such that the INT stimuli will be a reasonably representative sample of the various PROBS likelihoods.

The experimental design will include computation of normative solutions for all stages of all trials of all tasks. These solutions are required to ensure that all trials of all tasks are computable, i.e., fully-specified such that humans and models have all the information needed to compute a solution without further assumptions (e.g., from RASKRs). Normative solutions are also required to assess model predictions of human performance, as discussed further below.

4 Evaluation

Per the BAA, the evaluation of models will include three forms of assessment: Neural Fidelity Assessment (NFA), Cognitive Fidelity Assessment (CFA), and Comparative Performance Assessment (CPA).

4.1 Neural Fidelity Assessment (NFA)

Neural Fidelity Assessment (NFA) ensures that the models developed under the ICArUS program are grounded in neuroscience principles and consistent with the current understanding of neural circuits. The following is intended to amplify and supplement the guidance regarding this assessment that was provided in the Broad Agency Announcement (BAA, 2010).

Component Model Implementation Requirements

Table 4 below (reproduced from the BAA) lists the seven brain systems that ICArUS performers are expected to model. Additional brain systems may be incorporated, if needed, provided that they are deemed important for sensemaking and provided that the associated functionality (e.g. low-level vision) is not excluded by the BAA.

Neural Fidelity

An important goal of the ICArUS program is to create systems that can anticipate, prevent, or otherwise mitigate human error in intelligence analysis due to cognitive biases. The program’s neural fidelity requirements are important because cognitive biases are believed to arise from the brain’s underlying design features and resource limitations. Thus neural models, in comparison to purely functional (i.e. cognitive) models, are more likely to robustly predict human sensemaking strengths and weaknesses over a wide range of task parameters.

The architectures developed under the ICArUS program are not intended to represent complete models of the human brain or to incorporate complete models of specific brain regions. Rather, ICArUS models should replicate neural circuits that are necessary for supporting a broad range of sensemaking functions in a way that remains consistent with the current (although sometimes incomplete) understanding of neural structure and function.

Definitions

The BAA defines key neural fidelity concepts as they relate to the program:

> "*Biologically plausible* means that the model and its features are consistent with current neuroscientific understanding, even though that understanding may be incomplete. *Biologically implausible* means that the model (or some feature of the model) contradicts current scientific understanding of brain structure and function."

There are many ways to incorporate biological realism into a theory of sensemaking and various levels of detail at which neural circuits can be modeled. The required level of biological detail will depend on both the level of neuroscientific understanding of the corresponding brain system and on the elements needed to support the full range of sensemaking functions.

Table 4: Required brain systems and associated functions.

Brain system	Function
Prefrontal Cortex	Attention, cognitive control, working memory, goal-oriented behavior, decision making
Parietal Cortex	Evidence integration, decision making, multimodal sensory representation, spatial reasoning and memory, estimation of value and uncertainty
Temporal Cortex	Object representation, semantic knowledge representation
Medial Temporal Lobe, Hippocampus	Recognition and recall, declarative (episodic and semantic) memory, spatial cognition, relational processing, temporal sequence learning
Anterior Cingulate Cortex	Error signaling, cognitive control, conflict monitoring, decision making
Basal Ganglia / Dopaminergic Systems	Reinforcement learning, reward signaling, slow statistical learning, action sequencing, procedural learning, decision making
Brainstem Neuromodulatory Systems	Attentional arousal, transition between exploitative and exploratory behavioral modes

At a minimum, ICArUS models must not be biologically implausible. Models of cognitive processes that fail to comply with or contradict the known biology of the brain are out of scope and will result in a failing NFA score for the corresponding brain component. Where possible, component models must account for *how* their functions arise from their underlying neural circuitry. Hybrid models that combine symbolic and sub-symbolic elements are acceptable provided that such elements are convincingly linked to brain structure and function and that the overall approach remains strongly grounded in neuroscience. Purely functional elements are acceptable only when there is currently no understanding of the neural mechanisms responsible for performing those functions.

When multiple models of a brain component or function exist that are compatible with existing experimental data, performers are free to select the position that is most consistent with their overall theory of sensemaking provided that they justify their position with appropriate evidence and citations. If performers choose to implement simplified versions of existing models, they must explain how the simplified and more detailed models are functionally equivalent in terms of their sensemaking functions, cite published references describing the simplified equivalents, and justify why the removed subcomponents are unnecessary for sensemaking.

Requirements

According to the BAA, to exhibit neural fidelity, models of individual brain systems and their associated cognitive functions should:

1. Have a *structure* that is consistent with known neuroanatomical principles for the corresponding brain system.

2. Perform the same *cognitive function(s)* as the corresponding brain system.

3. Follow *internal dynamics* that are consistent with functional neuroimaging and electrophysiological studies.

4. Be *integrated* with other component models in a biologically plausible way based on known structural and functional connectivity found in the published neuroscience literature

Structure

Models should incorporate as much structural detail at or above the level of the neuron (e.g., excluding ion channels and precise dendritic morphology) as is reasonable and relevant to sensemaking. Component models must incorporate structures and pathways that are relevant for performing sensemaking tasks and predicting human performance (such as for incorporating cognitive biases) and, inversely, should not incorporate those that are irrelevant for these tasks. Structures and pathways within the models must be consistent with the known neuroanatomical principles for the corresponding brain systems. If an anatomical structure has major functional subdivisions (e.g., CA1 and CA3 in the hippocampus), then the associated component model should reflect this modular organization. The relative numbers of elements (i.e., neurons or neuron-like units) and element types found in the modeled brain areas should guide the characteristics of the associated component models.

Function

Component models must perform the same cognitive sensemaking functions as the brain systems they model. As with structure, functions that are irrelevant for sensemaking need not be incorporated into the models. Despite the program's emphasis on individual brain regions, we do not espouse the view (often found in popular media) that cognitive functions are neatly packaged in self-contained brain areas; as neuroscientists well understand, many of these functions are widely distributed across different regions and sometimes emerge from interactions between brain regions.

Purely computational, or algorithmic, models that create function without regard for the underlying structure (i.e., "black box models") are inconsistent with the program's objectives. Performers should justify how the computational approaches used in each model are consistent with published evidence.

Internal Dynamics

ICArUS models should reflect the current understanding of how component brain areas function and interact over time. The models' internal dynamics should match data from functional neuroimaging and electrophysiological studies that have characterized the temporal patterns of activity found in the brain. Concepts from control theory or dynamical systems theory may be helpful when describing these patterns. Where applicable, neuronal dynamics, including mean firing rates and/or precise spiking patterns, should be similar to those in real neurons.

Integration

The structural and functional networks connecting various brain regions should be represented faithfully in the overall architecture. Relevant neural pathways between connected areas should be present, and the connections between components should be consistent with physiological evidence. The BAA specifies the minimum number of component models that must be integrated at each waypoint/milestone, and all of the component models are expected to be fully integrated by Month 36 of the program.

Assessments

The BAA specifies requirements and general principles governing the Neural Fidelity Assessments. NFA evaluations will be qualitative in nature (pass/fail) and will be made by the ICArUS program manager with guidance from an independent panel of neuroscience experts with experience in computational neuroscience modeling, cognitive modeling, and artificial intelligence from government, academia, federally funded research and development centers, and university affiliated research centers. The panel's formal assessments will be reviewed independently by 1-2 government employees with expertise in computational cognitive neuroscience research.

Table 5: Metrics for each waypoint/milestone (Phases 1 and 2).

Month	Required Number of Implemented Component Models	Required Number of Integrated Component Models	Required Number of Component Models with "Pass" Score for NFA
Phase 1			
06 – June 2011	3	2	
11 – November 2011	4	3	
18 – June 2012	5	4	
23 – November 2012	5	5	3
Phase 2			
30 – June 2013	6	6	
36 – December 2013	7	7	
42 – June 2014	7	7	5

The Assessment Process

NFA evaluation is a continuous process based primarily on information available from performers' deliverables, including reports, software (code and executables), teleconferences, site visits and technical exchange meetings, and government/T&E team discussions. Additional information may be solicited from performers as needed. The NFA panel will report on performer progress and make recommendations to the program manager at each program waypoint/milestone (minimum 6-month intervals). Table 5 lists the NFA metrics associated with each program waypoint/milestone. During Phases 1-2, official assessments will be conducted only at months 23 and 42.

Within approximately a week following performer deliveries, the NFA panel will provide the program manager with preliminary evaluations of both the *fidelity* and *completeness* of individual component models based on their *anatomical* and *functional* properties. Biological plausibility was defined previously, and completeness is defined *relative* to the degree to which models incorporate the structures and functions from their associated brain areas considered necessary relative to the demands of sensemaking. The NFA panel will also indicate progress on component model integration and comment on the neural implementations of cognitive biases. A few weeks after the associated PI meeting (TEM), the NFA panel will also deliver a detailed report that assesses the neural fidelity of each performer's component models and integrated architecture. These evaluations will be qualitative (pass/fail) and will follow the requirements specified in the BAA as described in Table 2. At the program manager's discretion, information from the NFA evaluations may be shared with performers or may be incorporated as part of overall feedback.

Performer Guidance

The burden of proof is on performers to make a clear and convincing case for how and why their models meet the neural fidelity requirements using sufficient evidence (with citations) from the peer-reviewed neuroscience literature. Performers should be explicit regarding the key neurobiological features of each model, including learning algorithms and encoding schemes, the phenomena their models explain, and the levels at which their models represent the underlying anatomical structures and cognitive functions.

Reports and other deliverables should be prepared so that they facilitate understanding of the models they describe, highlighting both the strengths and limitations of the approaches used. Source code should be clearly documented and commented to a reasonable extent. Making the relevant information (i.e., claims and support) easy to find will help ensure that performers' accomplishments are represented accurately.

ICArUS performers were selected in part based on their combined diversity of modeling approaches, and it is understood that these different approaches create specific challenges for satisfying neural fidelity requirements. The differences between approaches will almost necessarily result in dissimilarities between the overall architectures, but all delivered systems must satisfy the program goals of incorporating biologically plausible neural models that perform sensemaking tasks while remaining faithful to known neural structures, functions, internal dynamics, and integration.

4.2 Cognitive Fidelity Assessment (CFA)

CFA (as well as CPA, discussed in Section 4.3) is concerned with how well a model predicts behavioral data on the challenge problem. More specifically, the goal of CFA is to assess whether models reproduce observed human cognitive biases. The following paragraphs describe quantitative measures of biases and how they are to be computed.

The primary measure of sensemaking response (see *Inference*, Section 2.1) is a distribution of probabilities across a set of hypotheses, e.g., $\{P_h\} = \{A\%, B\%, C\%, D\%\}$. These probability distributions for all subjects will be used to compute an average human response at each stage (j) of each trial (i) of each task, P_{ij}. Normative distributions Q_{ij} will also be computed and compared to P_{ij} using measures described below, as required for assessing cognitive biases in CFA. The corresponding model distribution M_{ij}, for comparison to human distribution P_{ij}, will be provided by each performer at each stage of each trial of each task.

An overall measure of uncertainty across a set of hypotheses, known as entropy (E), is computed as follows (Shannon & Weaver, 1949):

$$E_P = -\Sigma_h P_h * \log_2 P_h.$$

A normalized measure of negative entropy or "negentropy" (N) can be expressed on a scale of 0% to 100%, where N = 0% implies maximum uncertainty (i.e., maximum entropy) and N = 100% implies complete certainty (i.e., zero entropy). N is computed as follows:

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

where $E_{max} = 2$ for the case of four hypotheses (Tasks 2-6) and $E_{max} = 1$ for the case of two hypotheses (Task 1).

Negentropy enables comparison of how much certainty a human achieves in sensemaking to how much certainty he should have achieved. This in turn enables assessment of several cognitive biases. For example, if the distributions P and Q are such that $N_P > N_Q$ then the bias might be characterized as a *Confirmation Bias* – i.e., over-weighing evidence that confirms the most likely hypothesis. Conversely, if $N_P < N_Q$ then the bias might be characterized as the opposite of Confirmation Bias, known as *Conservatism* (Edwards, 1982), which can arise from various heuristics such as *Anchoring and Adjustment* (Tversky & Kahneman, 1974)

Table 6: Cognitive biases in six tasks of the challenge problem.

Bias[1,2]	**Task**					
	1	**2**	**3**	**4**	**5**	**6**
Anchoring and Adjustment					x	
Confirmation Bias						x
Representativeness	x	x	x			
Probability Matching				x	x	x

Notes:

1. For a given bias (row), a yes/no measure of the bias will be computed (using the corresponding metric in Table 7) on all trials of the tasks (columns) indicated by x. The total "yes" fraction across these trials will be used to assign a pass/fail score to a model in predicting the bias, i.e., "pass" = total "yes" fraction of at least 50%.

2. For each bias, trials will be excluded from the assessment if the bias exhibited (on average) by human subjects is not significant, as measured by a t-test applied to the corresponding metric in Table 7.

The four biases to be addressed in Phase 1, per BAA Table 3, are: *Anchoring and Adjustment*, *Confirmation Bias*, *Representativeness*, and *Probability Matching*. Anchoring and Adjustment, and Representativeness, are actually heuristics that lead to biases (Kahneman, et al., 1982), so CFA needs to specify the associated biases. As discussed further below, Anchoring and Adjustment as well as Representativeness lead to biases of *Conservatism*.

The challenge problem is designed such that each bias is expected to be exhibited by human subjects in one or more of the six tasks, as shown in Table 6. The biases are discussed further below, including a specification of how the existence of bias will be established computationally, e.g., by negentropy or some other parameter – see Table 7.

Table 7: Computational metrics to be used in assessing cognitive biases.

Bias	Metric[1,2]
Anchoring and Adjustment	$\lvert \Delta N_M \rvert < \lvert \Delta N_Q \rvert$
Confirmation Bias	$C > 0.5$
Representativeness	$N_M < N_Q$
Probability Matching	RMSE(F-P) < RMSE(F-I)

Notes:

1. N is negentropy. The subscripts M and Q are model and Bayesian, respectively. Refer to text for further details of notation.

2. A "pass" grade for a bias will be assigned to a model that satisfies the metric inequality for that bias. A "fail" grade for a bias will be assigned to a model that violates the metric inequality for that bias.

Anchoring and Adjustment

Anchoring and Adjustment (Tversky & Kahneman, 1974) is a heuristic by which an estimate is made (e.g., of a probability distribution) – starting from an initial value (i.e., an anchor) and then making adjustments to yield the final answer. The bias arises from insufficient adjustment, such that the final answer is biased toward the anchor. This heuristic and bias applies to the multi-stage inferences made by subjects in Task 5 of the challenge problem.

In the case of sensemaking, which requires estimation of probability distributions across a set of hypotheses, the anchor at each stage is a prior distribution that is either given to a subject or assumed by the subject (i.e., as the posterior from the prior update). The bias can be measured by negentropy, because inadequate (i.e., less than Bayesian) adjustment will cause $\lvert \Delta N_P \rvert < \lvert \Delta N_Q \rvert$. Typically this bias will produce *Conservatism*, which is characterized as $N_P < N_Q$ (see *Representativeness* below), because generally N_Q increases as more data are obtained and aggregated in stages of Task 5. The opposite behavior, in a form of Confirmation Bias, would be implied by $N_P > N_Q$. Pilot testing of human subjects clearly shows Conservatism and specifically Anchoring and Adjustment as defined by $\lvert \Delta N_P \rvert < \lvert \Delta N_Q \rvert$, on most trials of Task 5, hence this bias (and not a Confirmation Bias) is what models are required to predict in Task 5 (see Table 6).

Confirmation Bias

Confirmation Bias (Nickerson, 1998; Klayman & Ha, 1987; Fischhoff & Beyth-Marom, 1983) is a general term that refers to seeking or interpreting evidence in ways that are partial to existing beliefs. Here it is useful to distinguish between two types of Confirmation Bias: Confirmation Bias in Seeking Evidence; and Confirmation Bias in Weighing Evidence.

The process of interpreting evidence may include a wide range of effects that depend on the nature of the evidence and other context of the task. For AHA, evidence is presented in the form of INTS with unambiguous spatial features (not likely to be re-interpreted) and accompanied by PROBS with unambiguous numerical values (not likely to be misinterpreted). Thus, for our case the interpretation of evidence is a process by which numbers about features are selectively over-weighed or under-weighed as these numbers are aggregated. As implied by the phrase "partial to existing beliefs", a Confirmation Bias results from over-weighing evidence that supports the most likely hypothesis and under-weighing evidence that supports less likely hypotheses.

Based on pilot studies (mentioned above), the dominant behavior in Task 5 is Anchoring and Adjustment – which reflects a Conservative Bias rather than a Confirmation Bias in Weighing Evidence. Therefore the BAA bias of Confirmation Bias is assessed instead as a Confirmation Bias in Seeking Evidence, in Task 6 (see Table 6). Confirmation Bias in Seeking Evidence is a bias in decision-making, not a bias in sensemaking per se. However, the two are clearly related because inferences made in sensemaking provide the basis for choices made in decision-making. Moreover, when the decisions are choices about which evidence to seek, then decision-making in turn becomes a basis for inferences. This interaction is important because Confirmation Bias in Seeking Evidence has typically been tested in a deterministic context (Wason & Johnson-Laird, 1972) that clearly does not apply to the probabilistic nature of sensemaking.

In a deterministic task, a single piece of evidence can disprove a hypothesis and no amount of positive evidence can ever prove a hypothesis, so in that context attempts to confirm a hypothesis are non-normative. However, in a probabilistic context like that of sensemaking, no single datum can disprove or prove a hypothesis, because the problem is one of weighing various pieces of evidence in order to assess the relative likelihoods of various hypotheses. In that case, the heuristic known as a "positive test strategy" (Klayman & Ha, 1987), which produces an apparent Confirmation Bias in Seeking Evidence, is actually normative or at least approximately normative – in the sense that it maximizes the amount of expected information to be gained from seeking evidence. Thus, the so-called Confirmation Bias may not be a bias at all relative to optimal behavior, and instead the "bias" may be one in which a subject does not seek evidence that is expected to confirm the most likely hypothesis.

In the case of AHA, subjects seek evidence when they choose the INT type to receive next in stages of trials of Task 6. Each INT provides evidence about all hypotheses (A, B, C, D), not just the most likely hypothesis, so it is not clear how Confirmation Bias in Seeking Evidence could or should be measured for most INTS. However, for SIGINT a subject must choose the group on which to get SIGINT, and this choice of a group offers a clear way to test for a confirmatory strategy in the seeking of evidence.

With the parameters of AHA, normative calculations of Bayesian solutions show that the confirmatory strategy is actually optimal and hence not a bias in the sense of deviating from normative standards. Indeed, a subject would be biased in *not* always adopting the confirmatory strategy, so actually there can be no Confirmation Bias. Nevertheless, because a confirmation preference (aka positive test strategy) is often referred to as a Confirmation Bias, we adopt a definition of confirmation preference as the measure of Confirmation Bias required by the BAA. More specifically, we compute the fraction C of SIGINT choices for which the group G (on which SIGINT is requested) is the group with highest probability (as reported by the subject before receiving SIGINT). Pilot studies show that human subjects (on average) typically exhibit $C \approx 0.75$. Thus the threshold $C = 0.5$ serves as a pass/fail test for whether a model exhibits the bias (see Table 7). A model with $C > 0.5$ in Task 6 will be scored as exhibiting Confirmation Bias, and a model with $C \leq 0.5$ in Task 6 will be scored as not exhibiting Confirmation Bias.

Representativeness

In Representativeness (Tversky & Kahneman, 1974), a subject judges the probability that an object or event "x" belongs to a class "X" (or is generated by a cause "X") by the degree to which x is similar to X, based on some focal feature(s) typical (i.e., representative) of X. Like Anchoring and Adjustment, Representativeness is a general heuristic that can lead to various biases. One such bias in our case of geospatial sensemaking is “Regression to the Mean” – discussed in the original paper on heuristics and biases by Tversky & Kahneman (1974). Typically this bias resulting from Representativeness involves *misconceptions* of regression to the mean, for individual values *within* a distribution of values. However, in our case human subjects are reporting probability values *across* a probability distribution, and the measured behavior can be characterized as *exhibiting* regression to the mean. Presumably this is because human subjects consider regressed (flatter) probability distributions to be more representative than the actual distributions.

The resulting regression is a Conservative Bias, measured by $N_P < N_Q$, because the regressed probability distribution is flatter than the Bayesian probability distribution. Pilot studies show that on average human subjects exhibit this bias, as measured by $N_P < N_Q$, on Tasks 1, 2, and 3. Thus, these are tasks for which Representativeness will be assessed (see Table 6).

Probability Matching

Probability Matching is a tendency to select options at frequencies that are proportional to associated probabilities. This is a bias, relative to normative behavior, because expected utility is maximized by always (at 100% frequency) choosing the option with highest probability (assuming utility is the same for all outcomes, as it is in AHA where the consequences of attacks do not vary by group or by site).

Like Confirmation Bias in Seeking Evidence, Probability Matching involves a bias in decision-making rather than a bias in inferences (which provide the basis for decision-making). Also like Confirmation Bias in Seeking Evidence, Probability Matching is a reasonable (if not rational) behavior when one considers the ecological context in which decisions are typically made – especially the fact that likelihoods might change with time (Burns & Demaree, 2009). But unlike a confirmation preference in seeking evidence, Probability Matching is a non-normative behavior for tasks of the Phase 1 challenge problem where likelihoods are held constant in spatial context frames.

Probability Matching is often measured in simple experiments involving forced choices. However this requires that two conditions be met. First, there must be a large sample of choices by each subject, i.e., so that the frequencies of choices can be computed and compared to the subjects' probabilities reported in sensemaking. Second, the probabilities reported in sensemaking must be held constant (at least approximately) over the trials on which the frequency of choices is computed. Neither of these conditions is satisfied in the tasks posed by AHA. Thus Tasks 1-3, which involve forced choices, will not be used to assess Probability Matching.

Instead, Tasks 4-6 were designed to address Probability Matching in another context, involving allocation of resources in percentages against groups or sites. This provides a more direct measure of Probability Matching by avoiding the need to measure the frequency of repeated forced choices. Probability Matching will produce a Proportional Decision Bias, where the decision-maker's percentages in allocating resources tend to match the probabilities from sensemaking (reported immediately before making the decision). In contrast, a theoretically normative (non-biased) decision-maker would always allocate 100% of his resources against the group or site reported as having the highest probability of attack.

To measure Probability Matching in Tasks 4-6, the average human allocation of resources across groups F = {A%, B%, C%, D%} will be compared (on each trial) to two other distributions. One distribution is given by the average human probabilities P = {A%, B%, C%, D%). The other distribution is I = {100%, 0%, 0%, 0%}, where 100% is assigned to the group (A, B, C, D) with highest probability. The comparison will be quantified by a Root Mean Square Error (RMSE) on each trial, for both RMSE(F-P) and RMSE(F-I). Probability Matching produces RMSE(F-P) < RMSE(F-I), and this is indeed observed on Tasks 4-6 of pilot studies. Therefore, the equation RMSE(F-P) < RMSE(F-I) will be used in pass/fail fashion to assess whether a model predicts Probability Matching on trials of Tasks 4-6.

4.3 Comparative Performance Assessment (CPA)

CPA will assess a model's success in matching human performance, per the BAA Table 4 metric of a *50% success rate*. This requires definition of a relative success rate (RSR), which in turn is based on further measures described below.

A standard measure for comparing two probability distributions, like those of P (human) and M (model), is known as the Kullback-Leibler Divergence (KLD, Kullback & Leibler, 1951), discussed further in Appendix C. Consistent with the information-theoretic measure of entropy discussed above, K_{PM} measures the amount of information (in bits) by which the two distributions differ, computed as follows:

$$K_{PM} = E_{PM} – E_P = –\Sigma_h P_h * \log_2 M_h + \Sigma_h P_h * \log_2 P_h$$

where E_{PM} is the cross-entropy of P and M, and E_P is the entropy of P. Notice that $K_{PM} = 0$ when the distributions P and M are the same, and K_{PM} increases as M diverges from P.

KLD is a measure of divergence or "error", whereas the BAA requires a measure of "success". Also KLD ranges from zero (perfect match) to infinity (worst mismatch), whereas the BAA requires a measure of % success. To address these issues, we define a measure of similarity that ranges from 0% for the worst mismatch (when KLD is infinite) to 100% for a perfect match (when KLD is zero). This measure of similarity (S) is computed from K as follows:

$$S = 100\% * 2^{-K}.$$

As the divergence K ranges from zero to infinity, the similarity S ranges from 100% to 0%. Thus S_{PM} is useful for computing the success of a model M in matching human data P. However, S is an absolute measure (typically more than 50%, even for a poor match, because K is typically less than 1), so this measure of success must be further scaled relative to some standard for comparison to the BAA metric of 50%.

A common standard for assessing relative performance is a null model, which in our case would be a uniform distribution R = {0.25, 0.25, 0.25, 0.25}. This distribution has maximum entropy (minimum negentropy), and implies "random" (non-thinking) performance in sensemaking. Thus S_{PR} will be computed and used as the lower bound on S_{PM} in computing a relative success rate (RSR) as follows:

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

A model's RSR will be zero if S_{PM} is equal to or less than S_{PR}, because in that case a null model R would provide the same or better prediction of the human data P as the candidate model M. The RSR for a model M will increase as S_{PM} increases, up to a maximum RSR of 100% when S_{PM} = 100%. For example, if a candidate model M matches the data P with a similarity score of S_{PM} = 80%, and the null model R matches P with a similarity score of S_{PR} = 60%, then the RSR for model M would be 50%.

RSR will be computed at each stage of each trial of Tasks 1-5, which are the tasks on which all subjects receive the same stimuli. The average RSR across all trials of each task will be used to compute the performance of a model on the task. These task-average values of RSR on Tasks 1-5 will then be weighed, along with a metric for Task 6 (discussed below), in computing an overall score for CPA. RSR will not be computed for Task 6, because each subject receives different stimuli on each stage of each trial as the subject chooses INTS to receive (after receiving the initial INT from HUMINT). Instead a different metric is required to assess model predictions of human choices in Task 6. For this purpose the assessment will use a Relative Match Rate (RMR), as follows:

After receiving HUMINT at the start of each trial in Task 6, the subject's first choice is among four INTS (IMINT, MOVINT, SIGINT, or SOCINT). The next choice is among three remaining INTS, and the last choice is among two remaining INTS. Thus there are 4*3*2 = 24 possible sequences of choices that might be made by a subject on a given trial. For each trial, the fraction (%) of subjects that choose each of the 24 sequences will be computed. The sequence with maximum % will define a benchmark $F(t, s_{max})$ for each trial (t), where s_{max} refers to the sequence with maximum F for trial t. On the same trial, a model will predict a sequence s_{mod}, and the % value of $F(t, s_{mod})$ for this sequence will be computed from the human data. In words, $F(t, s_{mod})$ is the % of humans that chose the same sequence as the model chose, on a given trial t.

The Relative Match Rate (RMR) for INT choices on a trial (t) of Task 6 is then defined as follows:

$$\mathrm{RMR}(t) = F(t, s_{\mathrm{mod}}) / F(t, s_{\mathrm{max}}).$$

For example, assume a model predicts a sequence s_{mod} on a trial of Task 6. Assume also that 20% of human subjects chose the same sequence, but a different sequence was the most commonly chosen by human subjects, e.g., by 40% of subjects. In that case F(t, s_{mod}) = 20% and F(t, s_{max}) = 40%, so RMR(t) = 20 / 40 = 50%.

Finally, task-average scores for RSR on Tasks 1-5 will be combined with RMR on Task 6 using the following weighting factors:

Task 1 RSR:	0.05
Task 2 RSR:	0.10
Task 3 RSR:	0.15
Task 4 RSR:	0.15
Task 5 RSR:	0.40
Task 6 RMR:	0.15

Total CPA:	1.00

The average score for each task will be a simple average of all trials for that task. The above weights will then be applied to the task-average scores, with the following rationale for the weights: First, with six tasks being assessed in CPA, and each task addressing one or more of the sensemaking processes in BAA Table 1, a first cut would assign weight of 1/6 ≈ 0.15 to each task. This is appropriate for Task 4 RSR and Task 3 RSR (and for Task 6 RMR, discussed above). However, Task 2 and Task 1 are basically the same task, except that Task 2 involves four hypotheses rather than two hypotheses in Task 1. Therefore, Task 2 RSR will be assigned a weight of 0.10 and Task 1 RSR will be assigned a weight of 0.05, such that together the two weights sum to 0.15. The remaining weight of 0.40 will be assigned to Task 5 RSR. This is appropriate because Task 5 is of special importance, for three reasons, as follows:

First, Task 5 poses cognitive challenges of “all source” GEOINT most consistent with the notional challenge problem described in BAA Appendix E. Second, Task 5 tests humans and models on 40 responses (10 trials with 4 stages per trial), which is twice that of Task 4, four times that of Task 1, and eight times that of Task 2 or Task 3. Finally, the Task 5 challenge of multi-stage belief updating plays a key role in the Task 6 process of “acquire addition information” (i.e., acquiring information based on current beliefs, in order to update those beliefs). However, the nature of Task 6 precludes a robust measure of average performance in belief updating, because different subjects receive different stimuli based on the layer selections that they make. For all these reasons, Task 5 is considered especially important and will receive weight of 0.40 as described above.

Example calculations for RSR and RMR are provided in Appendix B. Further notes on KLD are provided in Appendix C.

4.4 Average Performance

For both CFA and CPA, a model is required to predict *average* human performance as discussed above. The model may do so by producing one "run", i.e., generating responses on each stage of each trial of each task, using fixed values for all model parameters. A model may instead produce multiple runs and compute the average of these runs, where values of model parameters vary between runs. But in that case the total number of runs (r) should be less than or equal to the total number of human subjects (n) that are tested to obtain average human responses, $r \leq n$. And, in either case, values for all model parameters must be fixed within each run for all stages of all trials of all tasks.

For both CFA and CPA, the assessment of each task includes only trials that are deemed to be significant. For CPA, significance is established by a chi squared test that compares the average human probability distribution (on each trial) to the null (uniform) distribution. Any trials of Tasks 1-5 that show no significant difference (measured by chi squared $p > 0.05$) will be excluded from the assessment (CPA) of neural models against human data.

For CFA, significance refers to the existence of a significant bias in human data relative to the normative standard. In this case, significance is established by a t-test that compares human response to the standard (normative) response as defined by the metrics in Table 7. For each bias, the assessment (CFA) only includes tasks for which significant bias is exhibited by human subjects on the majority of trials in the task. These tasks are noted in Table 6. For each bias in each task, specific trials are excluded if the bias is not significant (measured by t-test $p > 0.05$).

For those trials of a task that are significant, CFA will compute the fraction of trials on which a model predicts the bias – in accordance with the metrics of Table 7. Consistent with the 50% passing threshold for CPA, a model must predict the bias on ≥ 50% of the significant trials (across all tasks to which the bias applies) in order to receive a passing score for that bias on CFA. Consistent with BAA Table 4, a model must receive a passing score on at least two of the four biases in order to pass CFA for Phase 1.

5 References

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

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.

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.

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

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49(10),* 1295-1306.

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

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

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

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

MITRE (2012). *Geospatial Intelligence: A Cognitive Task Analysis. Part 1: Prescriptive Task Analysis.*

MITRE (2011). *A Computational Basis for ICArUS Challenge Problem Design*. Draft dated August 18, 2011.

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

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

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

Wason, P., & Johnson-Laird, P. (1972). *Psychology of Reasoning: Structure and Content*. Cambridge, MA: Harvard University Press.

Appendix A Task 7

A.1 Overview

The purpose of Task 7 is to investigate the phenomenon of *self-initiated* sensemaking. In Task 7, participants will decide *if and when* to seek additional data as well as which data (INT) to seek. In other words, participants will have the opportunity to *exploit* a particular hypothesis/frame for multiple consecutive trials up until the point when surprising feedback (prediction error) prompts them to *explore* alternative hypotheses/frames – that is, to begin the sensemaking process. As such, Task 7 embraces the notion of sensemaking as a state of heightened cognitive control relative to baseline.

To support self-initiated sensemaking, Task 7 differs from Tasks 1-6 in a few important ways. First, Task 7 employs a "wave-like" attack structure in which a particular group attacks for a variable number of contiguous trials before being replaced by a different group, which itself attacks for a variable number of contiguous trials, and so on. The participant is not told which group is attacking, but must infer the group from the observable data (SIGACTS) combined with other available data layers. The ability to infer the correct group (sensemaking) is necessary in order for the participant to succeed on the primary task of predicting the location of the next attack.

Another departure from Tasks 1-6 is that the underlying map (including scene objects such as roads and buildings) does not change from trial to trial. Preserving the scene setting across trials provides a fixed reference against which participants can compare the location of the current attack datum to previous attack data, thereby developing a cumulative picture of the evolving attack pattern.

A.2 Description

Self-initiated sensemaking. [15 trials]. Task 7 combines elements of Tasks 4 and 6 within a task structure that allows for *self-initiated* sensemaking.

Participants will be given four possible attack sites (1-4). As in Task 4, the primary task is to predict in which of several locations an attack will occur and to allocate resources to each location {@1%, @2%, @3%, @4%} – based on a normal distance decay function (HUMINT) and SOCINT – at the end of each trial. But unlike Task 4, to do so participants will have to have the right frame {P(A), P(B), P(C), P(D)}, i.e., an estimate of the group probability distribution.

To help, participants will be provided with a new rule:

> **If** a group attacked in the previous trial, **then** they are **85%** likely to attack again in the current trial.

A.3 Feedback

Deterministic feedback will be provided at the end of each trial (the location of each attack, represented by a dot in the SIGACT layer); the attacker (group) will be revealed once per wave, pseudo-randomly 1-4 trials before the end of a wave. The last 5 attack locations will persist on the screen. Based on that feedback, participants can decide if and when to seek additional information (IMINT, MOVINT, and SIGINT are available). Additional layer(s) can be "purchased" at the discretion of a participant (but are limited in number). Each layer will cost an associated number of credits (e.g., IMINT = 1; MOVINT = 1; SIGINT = 2).

Participants will start the task with 10 credits, and can earn additional credits through correctly predicting the location of attack. The number of earned credits will be calculated as follows: Assume that participants can earn a maximum of 1 credit in each trial. Now assume that on a single trial the participant assigns resources {40%, 30%, 20%, 10%} to locations {1, 2, 3, 4}, respectively. If the ground truth is location "2," then the participant would receive 0.3 credits (S2/100; see Section 2.3 for a description of S2 feedback).

The basic flow of the task is as follows:

1. At the beginning of each trial, participants will be asked to report the probability of attack by each group {P(A), P(B), P(C), P(D)} (similar to Tasks 5-6). Initial probabilities will be provided, based on HUMINT, in the first trial; participants will report changes *only when desired*.

2. Data (INTS) are presented, and participants are asked to estimate the probability of attack at each location {P(1), P(2), P(3), P(4)} (similar to Task 4).

3. Participants are asked to allocate troops to each location (% at 1, 2, 3, 4).

4. Participants are immediately shown *where* the attack took place (1-4). The attacker (A-D) will also be reported on pseudo-randomly chosen trials, once per wave.

5. Participants are given an opportunity to forage for additional information (INTS), limited only by the available number of credits.

6. The next trial begins.

A.4 Specification

As in Tasks 4-6, Task 7 involves rule-based sensemaking using data (INTS) and rules (PROBS). Table 3 in the main document specifies the rules for all INTS, including HUMINT, IMINT, MOVINT, SIGINT, and SOCINT. See Section 3 of the main document for additional details.

Appendix B RSR and RMR Calculation Example

B.1 RSR Example

For a trial with four hypotheses, assume the average human probabilities (in percent format) are **P (human) = [A: 22.54, B: 58.12, C: 6.43, D: 12.91]**, and that a model reports probabilities **M (model) = [A: 25.16, B: 53.52, C: 0.15, D: 21.17]**. We first apply a floor of 1% to any human or model probabilities that are below 1%. We then re-normalize the remaining values by adjusting each proportionally such that the resulting probabilities sum to 100%. In this case, the model has reported 0.15% for Group C, so we raise the probability of Group C to 1% and adjust the remaining values to obtain the following:

M =

[A: 25.16 - (0.85 * (25.16 / (25.16 + 53.52 + 21.17))),
B: 53.52 - (0.85 * (53.52 / (25.16 + 53.52 + 21.17))),
C: 0.15 + 0.85,
D: 21.17 - (0.85 * (21.17 / (25.16 + 53.52 + 21.17)))]

= **[A: 24.95, B: 53.06, C: 1.00, D: 20.99].**

We next compute S_{PM}, which is the similarity of the model (M) to the human data (P), as S_{PM} = 100% * 2^{-KPM}. Here K_{PM} is the Kullback-Leibler Divergence (K) computed as follows: $K_{PM} = -\Sigma_h P_h * \log_2 M_h + \Sigma_h P_h * \log_2 P_h$, where h is an index (h = 1-4) denoting elements in the P and M distributions. Note that probabilities should be in decimal and not percent format when computing K (e.g., 22.54% would be 0.2254). Thus we obtain:

S_{PM} = 100% * $2^{-0.1254}$ = **91.6726%.**

We next compute S_{PR}, the similarity of a uniform distribution **R = [A: 25, B:25, C:25, D:25]** to the data (P), as S_{PR} = 100% * 2^{-KPR}, where $K_{PR} = -\Sigma_h P_h * \log_2 R_h + \Sigma_h P_h * \log_2 P_h$. This yields:

S_{PR} = 100% * $2^{-0.4246}$ = **74.5023%.**

Finally, we compute RSR = 100% * (S_{PM} – S_{PR}) / (100% – S_{PR}), yielding:

RSR = 100% * (91.6726 - 74.5023) / (100 - 74.5023) = **67.34%**.

B.2 RMR Example

For a trial in Task 6, assume that the most frequently selected sequence by humans was “IMINT-MOVINT-SIGINT”, and that this sequence was selected at a frequency of 23% by humans. If a model also selects “IMINT-MOVINT-SIGINT”, then RMR for the trial would be computed as:

RMR = 100% * 23/23 = **100%.**

Suppose instead that a model selected the sequence “IMINT-MOVINT-SOCINT”, and that this sequence was selected at a frequency of 12% by humans. In that case, RMR would be computed as:

RMR = 100% * 12/23 = **52.17%.**

Appendix C Notes on KLD

KLD, used to compute S and RSR, is known to be asymmetric and nonlinear – and these features are both appropriate and advantageous for T&E. Distance metrics (linear and symmetric), such as Root Mean Squared Error (RMSE), might be defined based on the mathematics of geometry, but for ICArUS we are concerned with the mathematics of probability, and especially information theory. Consistent with information theory, KLD is computed as the sum over h (an index of probabilities in a discrete distribution) of $P_h * \log_2 (P_h/M_h)$, and both factors in this product are needed for properly comparing probability distributions between P and M.

First consider the factor $\log_2 (P_h/M_h)$. This nonlinear (log odds) term is an information-theoretic measure of "wow" (how much surprise, see Itti & Baldi, 2009) when a datum P_h is expected and the model M_h is reported. Notice that the wow factor is based on a ratio of P_h to M_h, which is a proportional measure rather than a linear difference. This scale-independent feature of KLD is a well-known advantage over scale-dependent metrics like RMSE or other distance measures. As a consequence, in KLD a given linear difference between P_h and M_h reflects a larger divergence at low probabilities, compared to the same linear difference at high probabilities. For example, when $P_h/M_h = 0.20/0.10$ then M_h is "off" (diverges from P_h) by a large proportion of P_h, whereas when $P_h/M_h = 0.90/0.80$ then M_h is "off" (diverges from P_h) by a small proportion of P_h, even though the linear difference between P_h and M_h is the same in each case.

Also in KLD, the factor P_h serves to weigh each wow in the sum, such that wows associated with high P_h are weighed more than wows associated with low P_h. That is, a high P_h in a distribution is more important than a low P_h in the same distribution, because hypothetical events with high P_h are more likely to occur than hypothetical events with low P_h. Therefore, each product $P_h * \log_2 (P_h/M_h)$ is a product of weight * wow, or importance * divergence, and the advantage of KLD lies in capturing the combination.

Because of KLD's scale-independence (discussed above), this metric is particularly sensitive to very low values of probabilities. Therefore, our assessment of RSR (based on KLD) will impose a "floor" that limits all human probabilities (averaged over subjects) and all model probabilities to a minimum value of 1%. Any value less than 1% will be changed to 1% and the remaining probabilities in the distribution will be adjusted accordingly such that all values are in the range 1-99%. This ensures that the data or models do not introduce artificial effects on KLD calculations at very low probabilities.

Appendix D Abbreviations

AHA	Abducting Hot-spots of Activity
BAA	Broad Agency Announcement
CFA	Cognitive Fidelity Assessment
COIN	Counterinsurgency
CPA	Comparative Performance Assessment
GEOINT	Geospatial Intelligence
HUMINT	Human Intelligence
IARPA	Intelligence Advanced Research Projects Activity
ICArUS	Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking
IMINT	Image Intelligence
INTS	Intelligence Sources
KLD	Kullback-Leibler Divergence
MOVINT	Movement Intelligence
NFA	Neural Fidelity Assessment
PROBS	Rule-based (if-then) Probabilities
RASKR	Rich and Sophisticated Knowledge Representation
RMR	Relative Match Rate
RMSE	Root Mean Square Error
RSR	Relative Success Rate
SIGACTS	Significant Activities Reports
SIGINT	Signal Intelligence
SOCINT	Socio-cultural Intelligence
T&E	Test and Evaluation