MTR140512
MITRE TECHNICAL REPORT

MITRE

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

Transition to the Intelligence Community

Kevin Burns

December, 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 14-4364

McLean, VA

Abstract

The IARPA (Intelligence Advanced Research Projects Activity) program ICArUS (Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking) developed and tested brain-based computational models of "sensemaking" – a cognitive component of intelligence analysis. MITRE's role was in Test and Evaluation (T&E) of the neural-computational models developed by several teams of performers. This document discusses the potential for transition of T&E products beyond ICArUS to practical applications and future research in the Intelligence Community. The products and potential uses are summarized by a table in Section 1 *Introduction*, and the most promising opportunities are highlighted in Section 3 *Conclusion.*

Table of Contents

1 Introduction

This document discusses potential uses of Test & Evaluation (T&E) products from ICArUS: Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking, a program funded by the Intelligence Advanced Research Projects Activity (IARPA). Details of the ICArUS program are provided in the Broad Agency Announcement (BAA, 2010) available at: http://www.iarpa.gov/index.php/research-programs/icarus/baa.

A summary of T&E products is provided in *ICArUS: Overview of Test and Evaluation Materials* (Burns, Bonaceto, Fine, & Oertel, 2014), available at: http://www.mitre.org/publications/all. The present document differs in addressing how these T&E products can be applied beyond the ICArUS program itself, to practical applications and further research in the Intelligence Community (IC).

As stated in the BAA (2010): *"The goal of the ICArUS Program is to construct brain-based computational models of the process known as sensemaking. Sensemaking, a core human cognitive ability, underlies intelligence analysts' ability to recognize and explain relationships among sparse and ambiguous data. By shedding light on the fundamental mechanisms of sensemaking, ICArUS models will enable the Intelligence Community to better predict human-related strengths and failure modes in the intelligence analysis process and will point to new strategies for enhancing analytic tools and methods."*

Consistent with that objective, the present document is concerned with how products of ICArUS T&E can *enable the Intelligence Community to better predict human-related strengths and failure modes in the intelligence analysis process and point to new strategies for enhancing analytic tools and methods.* The focus of this MITRE report is on T&E materials, rather than neural-computational models developed by ICArUS performer teams, for two reasons.

First, the role of MITRE and its subcontractors on the ICArUS program was to test and evaluate neural models developed by performer teams. Second, the scope of neural modeling in the program was limited to laboratory challenge problems designed by MITRE in accordance with T&E requirements of the BAA (2010). This scope excluded important aspects of real-world sensemaking that were deemed infeasible to address in integrated cognitive-neuroscience architectures, including natural language processing as well as the rich and sophisticated knowledge representations (RASKRs) by which intelligence analysts make sense of real-world situations.

With these scope limitations, the integrated cognitive-neuroscience architectures developed by ICArUS performer teams serve as laboratory demonstrations of computational models – rather than field-ready implementations of functional systems. Therefore the potential for near-term transition of ICArUS to practical applications lies primarily in the products of T&E, especially those products that extend beyond the program's scope of computational neural modeling.

The remainder of this report discusses eight such products and potential uses, as summarized in Table 1.

Table 1: Potential uses and users of ICArUS T&E products in the Intelligence Community (IC).

Product of T&E	**Purpose in T&E**	**Potential Uses in the IC**	**Potential Users in the IC**
Doctrinal Review	Defining the scope and methods of GEOINT.	Comparing the GEOINT perspectives of various agencies across the IC and DoD.	Analysts, Instructors, Engineers, Researchers
Analytical Stories	Identifying the challenges of sensemaking.	Training in the use of Bayesian analytic techniques.	Analysts, Instructors, Engineers, Researchers
Computational Basis	Modeling the components of analysis.	Developing techniques, training, and tools to support reasoning in accordance with Bayesian principles. Directing future R&D toward open-ended problems that require creative sensemaking.	Analysts, Instructors, Engineers, Researchers
Prototypical Problems	Exemplifying tasks of geospatial intelligence.	Relating the missions of ICArUS experiments to real-world challenges of geospatial intelligence.	Analysts, Instructors, Engineers, Researchers
Functional Software	Simulating geospatial intelligence missions.	Demonstrating the cognitive challenges of geospatial intelligence.	Engineers, Researchers
Mathematical Benchmarks	Evaluating sensemaking performance.	Quantifying the value of information and effectiveness of inferences.	Engineers, Researchers
Experimental Data	Uncovering biases and individual differences.	Understanding analytic performance and cognitive biases in sensemaking.	Engineers, Researchers
Behavioral Model	Computing human heuristics and biases.	Predicting cognitive biases as simplified deviations from normative standards.	Engineers, Researchers

As outlined in Table 1, the eight T&E products fall into two categories with respect to potential users identified in the rightmost column. The first four products are most generally applicable and potentially of use to analysts, instructors, engineers, and researchers. These products are discussed in Sections 2.1, 2.2, 2.3, and 2.4 below. The next four products are more specific to ICArUS and primarily of use to engineers and researchers. These products are discussed in Sections 2.5, 2.6, 2.7, and 2.8 below.

2 Discussion

2.1 Doctrinal Review: Defining the Scope and Methods of GEOINT

In accordance with T&E requirements of the ICArUS BAA (2010), MITRE designed "challenge problems" to measure the sensemaking performance of human beings and neural models. Also in accordance with the BAA (2010, page 10), these challenge problems involve *"… the analysis of simulated GEOINT data"*, where GEOINT is defined by the BAA (2010, page 8) consistent with Title 10 U.S. Section 467 as: *"The exploitation and analysis of imagery and geospatial information to describe, assess, and visually depict physical features and geographically referenced activities on Earth"*.

The program's focus on GEOINT was intended to promote transfer of insights from ICArUS R&D to stakeholders at the National Geospatial-Intelligence Agency (NGA), as well as other organizations in the IC that create or consume geospatial intelligence. As described in a recent report by the National Research Council (NRC, 2013) on the "*Future U.S. Workforce for Geospatial Intelligence"*:

> *"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 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."*

Sensemaking applies directly to this mission of NGA as an important aspect of *assessing the threats*. Thus as a first step toward design of ICArUS challenge problems, MITRE performed a review of geospatial intelligence doctrine – in order to better understand how various organizations and individuals defined the concept of GEOINT and prescribed the practice of GEOINT. This effort, known as the Prescriptive-Cognitive Task Analysis (PTA), was documented in a comprehensive report (MITRE, 2012a) and summarized in a companion paper submitted to NGA's journal, *Geospatial Intelligence Review* (Stech, 2014).

The PTA reviews doctrinal definitions of GEOINT put forth by NGA, DoD (Department of Defense, including Joint Staff, Marines, Army, and Air Force), and other organizations. The report also reviews various GEOINT methods, including: Intelligence Preparation of the Battlespace; GEOINT Preparation of the Environment; Structured Geospatial Analytic Method; and Querying, Mining, Assembly, Dissemination, and Integration. All of these methods are variations on a common theme in which functional processes such as *collection* (getting GEOINT), *depiction*, (showing GEOINT), *exploitation* (using GEOINT), and *dissemination* (sharing GEOINT) are integrated in the overall practice of geospatial intelligence.

Although details of the PTA were derived from documents published by government agencies, the MITRE report (2012a) and paper (Stech, 2014) hold potential for transition back to the IC by virtue of reviewing and relating the GEOINT perspectives of various agencies across the IC and DoD. Further potential for transition exists with respect to the *exploitation* of GEOINT, as ICArUS developed insights into how (and how well) analysts can use GEOINT to make sense of uncertain situations and thereby *"assess the threats"* (see NRC, 2013, noted above). For example, ICArUS formally identified the cognitive demands of GEOINT sensemaking, and experimentally uncovered cognitive biases such as *conservatism* that can compromise efficient exploitation of GEOINT by intelligence analysts. These and other insights from ICArUS, which are outlined in the remainder of this document, may be useful to *analysts* engaged in GEOINT; *instructors* who teach and train GEOINT; *engineers* who design systems that are employed in GEOINT; and *researchers* who advance the development of concepts, models, systems, or other aspects of GEOINT.

2.2 Analytical Stories: Identifying the Challenges of Sensemaking

The PTA (see Section 2.1) reviewed how GEOINT *should be* done per relevant doctrine. In addition, MITRE also reviewed how GEOINT *is done* by various analysts. This work is documented in a Descriptive-Cognitive Task Analysis (DTA), which includes 26 case-studies of geospatial sensemaking (MITRE, 2013).

The DTA is the product of structured interviews with intelligence analysts, who were encouraged to tell stories of cases in which they had to make sense of some anomalous information or uncertain situation. Analysts were also told that ICArUS was most interested in cases where a surprising discovery was made, while analyzing geospatial information, which enabled the explanation or prediction of some entity or activity. Over a dozen interviews were performed with analysts across various agencies of the IC and DoD, and these interviews were supplemented by reviews of published papers addressing similar problems of geospatial analysis.

The primary purpose of the DTA was to ensure that ICArUS challenge problems could be related directly to GEOINT practice as well as to GEOINT doctrine (per the PTA). An examination (see Burns, 2014b) of all 26 stories in the DTA uncovered six inferential variables that were repeatedly involved in sensemaking about adversarial situations (e.g., assessing the threats). The six variables, identified in Table 2 for each of the 26 stories in which they appear, are characterized as follows: *vulnerability*, *opportunity*, *capability*, *activity (prognostic)*, *activity (forensic)*, and *propensity*. These variables are routinely addressed in analytic practices such as "Suitability Analysis", which includes assessments of *vulnerability*, *opportunity*, and *capability*; and "Activity-Based Intelligence", which involves prognostic and forensic inferences about hypothesized *activities* as well as associated *propensities* of adversaries to act. Therefore the ICArUS Phase 2 challenge problem (Burns, 2014b) was designed to include all six variables in diverse "missions" posed by ICArUS experiments. The ICArUS Phase 1 challenge problem (Burns, Greenwald, & Fine, 2014) involved similar variables, but focused on spatial variations without temporal dependencies.

Table 2: Mapping six variables of the ICArUS challenge problems to 26 case studies of intelligence. The variables are as follows: P = vulnerability, U = opportunity, P_c = capability, P_t = activity (prognostic), F_t = activity (forensic), and P_a = propensity. (For details see Burns, 2014b).

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	

Besides the primary purpose of guiding challenge problem design, a secondary purpose of the DTA was to promote a program-long dialogue between ICArUS researchers and IC stakeholders. The DTA stories helped ICArUS researchers understand the cognitive demands and context of real-world sensemaking. The mapping of these stories to ICArUS challenge problems helped IC stakeholders understand how laboratory research on sensemaking relates to real-world analytic practice.

Consistent with this secondary purpose, DTA may be useful for continuing transition of ICArUS insights to the IC. As a specific example, MITRE has identified the potential for developing a new analytic technique to support Bayesian reasoning, as discussed below in Section 2.3. The DTA stories can serve as real-world case studies for training analysts in use of this technique, and for testing analysts on their understanding of the underlying principles.

2.3 Computational Basis: Modeling the Components of Analysis

Along with a focus on GEOINT (see PTA and DTA in Sections 2.1 and 2.2), the BAA (2010) required that ICArUS challenge problems address core sensemaking processes outlined in a so-called *"data-frame theory of sensemaking"* (Klein, et al., 2007).

The data-frame theory offers a conceptual description of sensemaking, but does not provide a computational specification of functional processes or knowledge representations – as needed for rigorous design and assessment of ICArUS challenge problems. In particular, the ICArUS BAA (2010) required that T&E include two types of quantitative assessments, namely: Comparative Performance Assessment (CPA), using a numerical percentage to measure how well a neural model matches human sensemaking performance; and Cognitive Fidelity Assessment (CFA), using normative (Bayesian) solutions as benchmarks for measuring whether neural models exhibit cognitive biases like those of human subjects.

To meet these requirements, MITRE developed a Bayesian framework that models sensemaking in a recurring cycle of eight stages, dubbed the *Octaloop* (Burns, 2014a). The stages are numbered and named as follows:

1. *Isolate Evidence*
2. *Generate Hypotheses*
3. *Estimate Likelihoods*
4. *Aggregate Confidence*
5. *Prognosticate Consequence*
6. *Evaluate Consequence*
7. *Anticipate Evidence*
8. *Discriminate Evidence*.

Steps 1-4 model processes of *inferencing*. Steps 5-6 model processes of *decision-making*, which rely on results from inferencing. Steps 7-8 model processes of *foraging*, which rely on results from inferencing and then support inferencing in the next cycle of sensemaking.

This model, derived from a real-world story of sensemaking by Klein, et al. (2007), was used to design challenge problems that satisfied two major program constraints. First, the challenge problems were designed to pose cognitive demands of geospatial intelligence, including core sensemaking processes identified in the BAA (2010). Second, the challenge problems were designed to enable quantitative assessments of human and model performance, per the BAA (2010), including the computation of normative (Bayesian) solutions needed for measuring cognitive biases exhibited by human subjects and neural models in ICArUS experiments.

In light of other program constraints, discussed in Section 1 (also see Burns, 2014a), the ICArUS challenge problems are necessarily simplified with respect to real-world sensemaking. The Octaloop, which was derived from examples of real-world sensemaking, helps pinpoint exactly what cognitive representations and processes are simplified in ICArUS challenge problems, and why the simplifications were necessary in order to accomplish T&E per the ICArUS BAA (2010). The Octaloop also helps point to how these simplifications can be overcome in transition to IC applications beyond the scope of ICArUS itself. In particular, Section 7 *Transition* of the Octaloop document (Burns, 2014a) discusses practical applications to analytic *techniques*, *training*, and *tools* – as well as to further *R&D*.

With respect to *techniques*, Section 7.1 of the document (Burns, 2014a) suggests that the Octaloop can be used to develop a new Structured Analytic Technique (SAT) – dubbed HELP (hypotheses, evidence, likelihoods, priors, and posteriors). This technique addresses limitations of existing SATs (Beebe & Pherson, 2012), such as Brainstorming and Analysis of Competing Hypotheses (ACH, see Heuer, 1999), which do not support analytical reasoning in accordance with Bayesian principles. Initial transition of HELP to intelligence practitioners and researchers has been accomplished by a paper submitted to the *International Conference on Naturalistic Decision Making* (Burns, 2014c).

With respect to *training*, Section 7.2 of the Octaloop document (Burns, 2014a) suggests that sensemaking stories can be useful for training HELP as a Bayesian SAT. Besides the story by Klein, et al. (2007), which was used to develop and demonstrate the Octaloop, all 26 stories of sensemaking contained in the DTA (MITRE, 2013) have been informally reviewed to ensure that the Octaloop and associated HELP technique do apply. A more formal application of Bayesian HELP to any of these stories could be performed in order to support case-based training – i.e., to tailor the training to case studies that are most relevant to an individual analyst's interests and expertise.

With respect to *tools*, Section 7.3 of the Octaloop document (Burns, 2014a) highlights those steps of cognitive processing that were performed for human participants (and neural models) in ICArUS experiments, rather than by participants in these experiments. For example, the four Octaloop steps of *Discriminate Evidence*, *Isolate Evidence*, *Generate Hypotheses*, and *Estimate Likelihoods* were all performed by the computer. The results of these steps were then provided to participants as inputs from simulated intelligence systems, which participants used in the Octaloop step of *Aggregate Confidence* (i.e., to evaluate the probabilities of competing hypotheses given evidence from various sources/systems). These simulated intelligence systems represent prototype *tools* that might be extended and implemented in real-world systems, to support human sensemaking by automating the associated steps of the Octaloop.

Those steps of the Octaloop that were performed by participants in ICArUS experiments include *Aggregate Confidence*, *Prognosticate Consequence*, *Evaluate Consequence*, and *Anticipate Evidence*. At these steps ICArUS measured and modeled human sensemaking, including cognitive biases relative to normative (Bayesian) solutions. The experimental results suggest a number of opportunities for designing tools that can help humans overcome biased behaviors – i.e., by automating aspects of the Bayesian SAT dubbed HELP. For example, a strong bias observed in the step of *Aggregate Confidence* was conservatism, in which participants reported probability distributions that were too close to {0.50, 0.50} compared to the Bayesian solution for two hypotheses {H, ~H}. Computational systems might be developed to accomplish this step and/or advise analysts on how to aggregate more effectively. One example of such a system uses "structure-mapping" visualizations to assist users in applying Bayesian principles (Burns, 2006; 2007), and this as well as other opportunities for support systems are discussed in the Octaloop document (Burns, 2014a).

With respect to *R&D*, Section 7.4 of the Octaloop document (Burns, 2014a) suggests that a Bayesian approach can help bridge the gap that currently exists between the practice of intelligence analysis and research efforts in academia and industry. In particular, the Octaloop document identifies those cognitive-computational processes that are necessary for competence in sensemaking, highlighting which processes have been addressed directly by ICArUS R&D and which processes remain to be addressed in future R&D. As noted above, the greatest opportunities for further research appear to exist at the Octaloop steps of *Discriminate Evidence*, *Isolate Evidence*, *Generate Hypotheses*, and *Estimate Likelihoods*, which were all performed for participants (by simulated systems) rather than by participants in experiments.

As discussed in the Octaloop document (Burns, 2014a), these cognitive processes could not be measured and modeled by challenge problems within the scope of the ICArUS program, for several reasons. One reason is that the BAA (2010) required measuring and modeling of *average* human performance, which meant that all human participants had to use the same evidence, hypotheses, and likelihoods. Another reason is that visual perception and natural language processing were *out of scope* for neural modeling and hence not tested in ICArUS challenge problems. These two cognitive capabilities are especially important in the Octaloop steps of *Discriminate Evidence* and *Isolate Evidence*. Finally, the BAA (2010) required minimizing any role played by humans' RASKRs, which are the cognitive basis by which humans generate hypotheses and estimate the likelihoods of evidence (in light of hypotheses). Minimizing the role of RASKRs meant eliminating the steps of *Generate Hypotheses* and *Estimate Likelihoods* from the scope of cognitive demands posed by ICArUS challenge problems.

Clearly these cognitive processes and the RASKRs that they employ are important to real-world sensemaking. In fact, the steps of *Generate Hypotheses* and *Estimate Likelihoods* are arguably the most important steps of all – as they provide the foundation for subsequent steps in which analysts make sense of evidence by assessing the probabilities of competing hypotheses. Therefore one important direction for future research is to address "creative" sensemaking – in "open-ended" problems where hypotheses, likelihoods, and evidence are not pre-defined by experimenters and not provided to participants as inputs.

2.4 Prototypical Problems: Exemplifying Tasks of Geospatial Intelligence

As described above, efforts to design ICArUS challenge problems were informed by GEOINT doctrine and practice as well as by a Bayesian-computational model of sensemaking. The resulting challenge problems pose cognitive demands that are prototypical of geospatial intelligence analysis and are quantifiable in accordance with the ICArUS BAA (2010).

Different challenge problems were developed for Phases 1 and 2 of the program, with Phase 1 focused on spatial sensemaking and Phase 2 focused on spatial-temporal sensemaking. Detailed documents describe the design rationale and associated test specifications for Phase 1 (Burns, Greenwald, & Fine, 2014) and Phase 2 (Burns, 2014b).

The test specifications are necessarily specific to the ICArUS program itself. Therefore the underlying design rationale is likely to be among the most useful aspects of ICArUS challenge problems in transition to IC applications. In particular, the Phase 2 design document (Burns, 2014b) includes a Section 6 on *Transition* that maps all 26 case studies of sensemaking obtained from DTA (see Section 2.2) to key variables of the challenge problem. The same mapping also addresses practical applications to tools, training, and techniques, similar to the discussion summarized in Section 2.3 above.

Besides the challenge problems themselves, associated products of ICArUS T&E include *Functional Software*, *Mathematical Benchmarks*, *Experimental Data*, and a *Behavioral Model*. All of these products might be leveraged in engineering applications and future research investigations, as discussed in Sections 2.5, 2.6, 2.7, and 2.8 below.

2.5 Functional Software: Simulating Geospatial Intelligence Missions

A useful feature of the ICArUS challenge problems is that they are implemented in functional software – as employed in ICArUS experiments. This software might be adapted for other uses, for example to assess the cognitive biases of individual analysts, and for training analysts to overcome their biases. However, such applications would require major modifications to the ICArUS challenge problem software, for several reasons.

First, although the challenge problem software measures individual responses, this was done per the BAA (2010) in order to obtain a robust *average* of performance and biases across a population of participants (roughly N = 100 in each phase of the program). The experimental design of stimuli and measures of performance did not address matters of consistency within each individual's responses, as would be needed to establish the reliability of an instrument for testing an individual's biases.

Second, the challenge problem software was designed for *testing* human subjects (and neural models), and for *measuring* cognitive biases relative to normative standards. The software contains no functionality for *training* human subjects or otherwise *mitigating* cognitive biases. In fact, the software was specifically designed not to introduce any training that would mitigate

biases, because the purpose of ICArUS experiments was to measure biases as they exist naturally. A major re-design effort would be required to augment the existing software with capabilities for training analysts and mitigating biases.

Because the ICArUS challenge problem software was designed for different purposes, it is not recommended for use in testing individual analysts or training analysts to overcome biases. Instead the computational basis for design of ICArUS challenge problems, namely the *Octaloop*, is recommended for these uses as discussed in Section 2.3. Nevertheless, ICArUS challenge problem software may be useful for *demonstrating* the cognitive task demands of geospatial sensemaking. This is because the software was designed to engage participants in game-like simulations of geospatial intelligence "missions", prototypical of real-world analysis (within program constraints), and great care was taken by designers in communicating the missions to participants as needed to conduct meaningful experiments. In particular, the software for each challenge problem includes an overall tutorial and mission-specific instructions, presented to participants at the start of an experiment and available for reference throughout the experiment. These materials are non-technical and understandable by experts and non-experts in disciplines of geospatial analysis, as evidenced by successful completion of all missions by hundreds of diverse participants in ICArUS experiments.

As a product of T&E, all screen shots comprising the tutorials and mission instructions are captured in a "walkthrough" document for Phase 1 (Burns, Fine, Bonaceto, & Beltz, 2014) and Phase 2 (Burns & Bonaceto, 2014). Compared to the challenge problem design documents (Burns, Greenwald, & Fine, 2014; Burns, 2014b), the walkthrough documents provide a much more accessible introduction to ICArUS challenge problems. Therefore these walkthrough documents would be a good starting point for those interested in using the challenge problem software for purposes beyond ICArUS experiments. For software developers, further details of the input/output file formats are available in a development guide for Phase 1 (Bonaceto & Fine, 2014a) and Phase 2 (Bonaceto & Fine, 2014b). Other files and folders included in the challenge problem software are described by *ICArUS: Overview of T&E Materials* (Burns, Bonaceto, Fine, & Oertel, 2014).

2.6 Mathematical Benchmarks: Evaluating Sensemaking Performance

Across both phases of the ICArUS program, an important aspect of challenge problem design was to compute normative (Bayesian) solutions – as benchmarks for measuring the cognitive biases of human subjects and neural models. This was a difficult requirement to satisfy, as discussed in *A Computational Basis for ICArUS Challenge Problem Design* (Burns, 2014a). As a result, the specific task demands of Phase 1 and Phase 2 challenge problems were shaped largely by the cognitive biases (relative to normative solutions) that were specified by the BAA (2010) for Cognitive Fidelity Assessment (CFA). There were four biases in Phase 1, and four more for a total of eight biases in Phase 2 as outlined in Table 3.

Table 3: Cognitive biases addressed in ICArUS challenge problems.

Behavior (bias)	**Example (referring to evidence, hypothesis, and confidence)**
Anchoring and Adjustment	When new evidence supports a hypothesis, a person's confidence in that hypothesis goes up less than it should.
Persistence of Discredited Evidence	When new evidence refutes a hypothesis, a person's confidence in that hypothesis goes down less than it should.
Representativeness	When evidence is typical of a hypothesis, a person discounts other relevant evidence (such as base rates) in assigning confidence to the hypothesis.
Availability	When a hypothesis (or evidence supporting a hypothesis) is vivid or otherwise memorable, a person assigns too much confidence to the hypothesis.
Probability Matching	When given a choice among options, a person will not always choose the option that is probably the best, but rather will choose each option at a frequency equal to its probability of being the best.
Confirmation Bias	When given a choice among sources of evidence, a person will seek evidence that is likely to support the most probable hypothesis.
Satisfaction of Search	When collecting evidence to evaluate a hypothesis, a person will prematurely terminate the search upon finding supporting evidence.
Change Blindness	When monitoring evidence to evaluate a hypothesis, a person will fail to detect evidence that supports another hypothesis.

In order to assess these biases, design of the challenge problems (Burns, Greenwald, & Fine, 2104; Burns, 2014b) included development of a mathematical equation (inequality) that defined each bias as a deviation from normative (Bayesian) performance. These defining equations were used in Cognitive Fidelity Assessment (CFA), and they are also useful beyond ICArUS as concise and computable statements of the biased behaviors. Often in the IC and elsewhere, claims about biases are lacking normative benchmarks and objective measures as needed to establish the existence and extent of biases. The defining equations developed for use in CFA offer a more rigorous approach, at least for the biases and task demands within the scope of the ICArUS program. These defining equations may be useful in future research and practice aimed at uncovering and overcoming biases in intelligence analysis.

Similarly, Comparative Performance Assessment (CPA) required mathematical measures of sensemaking performance – in order to compare human and model performance to each other as well as to normative solutions. These mathematical measures (Burns, 2014b; Burns, Greenwald, & Fine, 2014; Burns, 2011) employ information-theoretic concepts, such as relative entropy, to quantify the overall divergence of one belief structure (e.g., human or model) from another (e.g., Bayesian or random). The same information-theoretic measures were used to assess the "value of information" – i.e., to measure how strongly some evidential data (from intelligence sources) should affect the beliefs of a human or model. Like the measures of cognitive biases discussed above for CFA, these measures of sensemaking performance may be useful in future research as well as practical applications – to quantify the value of information and effectiveness of inferences under uncertainty.

Beyond ICArUS, the mathematical measures developed for CFA and CPA would likely be of most use to engineers who design advanced systems for "fusing" intelligence information. In particular, the state of the art in "high-level information fusion" is known to be lacking models and measures of sensemaking, as discussed by Blasch, et al. (2012). The Bayesian-mathematical Octaloop (Burns, 2014a) and associated information-theoretic measures of performance (Burns, 2014b; Burns, Greenwald, & Fine, 2014; Burns, 2011) can advance this state of the art – by providing a formal model and measures to quantify human, machine, and combined human-machine performance. Previously such efforts have been hampered by a lack of theories and models that express human sensemaking in computational terms, consistent with those used in the engineering of machine systems and needed for integrating human-machine performance.

Also within the realm of data fusion, a notion of "layers" or "levels" appears in almost all work aimed at integrating systems and humans. Although obviously simplified, the ICArUS challenge problems are useful for understanding how levels of human sensemaking relate to layers of system designers. In particular, besides characterizing *stages* of sensemaking in the Octaloop, the ICArUS challenge problems also highlight causal-hierarchical *levels* at which information is processed in sensemaking, as follows: intent → tactic → action → feature → datum; where arrows indicate the direction of causality (i.e., from cause to effect: cause → effect). This hierarchy was made explicit in design of the Phase 2 challenge problem, and the Phase 2 design document (Burns, 2014b) provides a brief discussion of how the five levels relate to five layers of a well-known data fusion model (Steinberg & Bowman, 2004).

2.7 Experimental Data: Uncovering Biases and Individual Differences

Besides the functional software and mathematical benchmarks used for ICArUS experiments, MITRE has also made available the human behavioral data collected in Phase 1 (N = 103 participants) and Phase 2 (N = 123 participants) of the program. The software/data package includes source code for analyzing all behavioral data, along with results of data analyses in numerical and graphical formats, as described in the T&E *Overview* document (Burns, Bonaceto, Fine, & Oertel, 2014).

These human data may be useful to engineers and researchers interested in measuring and modeling cognitive performance. As described in Section 2.6, the challenge problem missions were designed to elicit and evaluate eight different cognitive biases, each defined by a mathematical equation (inequality) based on information-theoretic quantities such as entropy. ICArUS experiments involve a diverse set of missions and biases, which offer a rich dataset to researchers who wish to model cognitive performance on laboratory tasks that are typical of geospatial intelligence. This was the purpose for which human data were collected in ICArUS experiments, i.e., for use by researchers building neural-computational models, and the same data may be useful to other researchers in future modeling efforts.

Per the BAA (2010), ICArUS experiments were aimed at measuring and modeling *average* human performance across all participants in each experiment. However, some research teams wished to explore the potential for predicting *individual* human performance. This was of interest to ICArUS from a transition perspective, because of its potential for informing personnel selection in the practice of real-world intelligence. One hypothesis was that individual differences could be attributed to cognitive "phenotypes" and distinguished by psychometric instruments based on underlying neural attributes – such as the scales for Behavioral Inhibition System (BIS) and Behavioral Activation System (BAS) proposed by Carver and White (1994). Also a number of demographic variables were hypothesized to affect individual differences in performance, including years of experience as a practicing analyst, and level of training in probability and statistics.

MITRE performed regression analyses (for numerical variables) and Analyses of Variance (ANOVAs, for categorical variables) to establish whether any of these or other independent variables affected dependent measures of performance in the Phase 2 experiment ($N = 123$). There were three dependent measures of interest, namely: *inferencing* (about hypotheses given evidence), *decision-making* (about the best course of action, given results of inferencing), and *foraging* (to obtain information for use in inferencing and decision-making). Results showed that none of the dozen-plus psychometric or demographic variables that were analyzed could significantly ($p < 0.05$) and substantially ($R^2 > 10\%$) predict any of the dependent measures of performance (inferencing, decision-making, or foraging). These results suggest that individual differences in sensemaking performance are very difficult to predict, and that such differences are not readily attributed to parameter variations in neural models. The results also support the approach of the ICArUS BAA (2010), which was to focus the program on measuring and modeling *average* sensemaking performance (rather than individual differences) in order to understand the existence and extent of cognitive biases relative to normative standards.

2.8 Behavioral Model: Computing Human Heuristics and Biases

As a final T&E product, MITRE developed a behavioral model for *Computing Analytical Biases*, dubbed *CAB*. This model differs from the Bayesian Octaloop (Section 2.3), which is a normative model of the sensemaking cycle. CAB is a cognitive model, using simple equations to compute human sensemaking responses and to characterize biases as deviations from Bayesian computations – for specific tasks/trials that were presented to participants in ICArUS experiments.

For instance, some tasks of ICArUS challenge problems presented probability distributions that participants were required to aggregate in order to evaluate competing hypothesis. The normative procedure for doing so, per the Octaloop step of *Aggregate Confidence*, is to compute the normalized product of input probabilities. As an example, if P_1 = {0.80, 0.20} and P_2 = {0.60, 0.40} are provided as independent probability distributions for hypotheses {A, B}, then the Bayesian aggregation is P' = {0.86, 0.14}, which is computed as a product of P_1 and P_2 normalized across the two hypotheses as follows:

$$P'(A) = (0.80 * 0.60) / [(0.80 * 0.60) + (0.20 * 0.40)] = 0.86$$

$$P'(B) = (0.20 * 0.40) / [(0.80 * 0.60) + (0.20 * 0.40)] = 0.14$$

As modeled by CAB, the cognitive heuristic for combining P_1 and P_2 is to average the two probability distributions, which produces P' = {0.70, 0.30}. Notice that this result is conservative (also called "regressive") relative to the Bayesian result P' = {0.86, 0.14}, because P' = {0.70, 0.30} is closer to a maximum entropy distribution {0.50, 0.50}.

This conservative bias was observed throughout ICArUS experiments (as well as other experiments that have been published, see Edwards, 1982), and the magnitude of bias was found to be well-predicted by the averaging heuristic discussed above. Thus for the cognitive task of aggregating two independent probability distributions (e.g., P_1 and P_2) regarding the same set of hypotheses (e.g., {A, B}), the CAB model uses a simple average instead of computing the Bayesian-normalized product. Similarly, CAB implements other simple equations to model other cognitive biases as heuristic deviations from Bayesian computations.

The simple equations of the CAB model were found to be remarkably accurate in predicting human performance and sensemaking biases (MITRE, 2014; 2012b). In fact the CAB model was comparable to the best neural models developed by ICArUS performer teams in predicting human responses and biases, per the program metrics for CPA and CFA (discussed in Section 2.6).

Of course CAB is only a parametric model of heuristic strategies, and as such does not offer mechanistic insights into underlying neural-biological processes. But computational simplicity and behavioral accuracy combine to make CAB a useful model for predicting the magnitudes of analytical biases in laboratory situations as well as real-world intelligence.

For example, as discussed in Section 2.3, training of a Bayesian technique dubbed HELP has been proposed as one way to help analysts overcome the conservative bias found throughout ICArUS experiments. One use of the CAB model would be to compute the expected magnitude of bias in practical situations, as a measure of the potential benefit to be gained by the Bayesian technique. In addition to the *averaging* heuristic discussed above, CAB also computes other heuristics and biases at other steps of the Octaloop. This makes the model useful for assessing which steps of the Octaloop, and associated heuristics/biases, would be most beneficial to address with new analytic techniques, training, and tools.

3 Conclusion

This document reviewed eight products of ICArUS T&E, and discussed how each might be extended beyond ICArUS to practical applications and future research in the Intelligence Community (IC). Referring to Table 1 of Section 1 *Introduction*, the eight products fall into two categories. The first four products deal most generally with the cognitive challenges of sensemaking, as prescribed by geospatial intelligence doctrine (Section 2.1), described in analytical case studies (Section 2.2), modeled in a Bayesian-computational framework (Section 2.3), and captured in the experimental designs of ICArUS challenge problems (Section 2.4). These products hold promise for use by analysts, instructors, engineers, and researchers across the IC. The next four products deal more specifically with software (Section 2.5), metrics (Section 2.6), data (Section 2.7), and insights (Section 2.8) from ICArUS experiments and analyses. These products would be most useful to information system engineers and researchers developing computational-cognitive models.

Among the eight products, one appears especially promising for near-term transition to the IC. This product is the Bayesian-computational *Octaloop* that was developed as *A Computational Basis for ICArUS Challenge Problem Design* (Burns, 2014a). As detailed in Section 7 *Transition* of that document, and discussed in Section 2.3 of the present document, the Octaloop can be used as a Structured Analytic Technique (SAT) to support inferencing in accordance with Bayesian principles. This technique, dubbed HELP (hypotheses, evidence, likelihoods, priors, and posteriors), could be taught in the context of real-world case studies such as those used to develop the Octaloop in the first place. The technique and associated training are discussed further in a paper submitted to the *International Conference on Naturalistic Decision Making* (Burns, 2014c).

With respect to future R&D, an important direction is to measure and model "creative" sensemaking on "open-ended" problems that more directly address the Octaloop steps of *Discriminate Evidence*, *Isolate Evidence*, *Generate Hypotheses*, and *Estimate Likelihoods*. Human performance on these aspects of sensemaking could not feasibly be measured and modeled within the constraints of the ICArUS program. Nevertheless, these aspects of sensemaking are obviously important and arguably of most importance to real-world problems of intelligence analysis.

4 References

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

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

Blasch, E., Jousselme, P., Lambert, D., & Bossé, É. (2012). Top ten trends in high-level information fusion. *Paper Presented at the International Conference on Information Fusion*.

Bonaceto, C., & Fine, M. (2014a). *Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking (ICArUS): Phase 1 Test and Evaluation Development Guide*. MITRE Technical Report, MTR130652.

Bonaceto, C., & Fine, M. (2014b). *Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking (ICArUS): Phase 2 Test and Evaluation Development Guide*. MITRE Technical Report, MTR140472.

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

Burns, K. (2014b). *Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking (ICArUS): Phase 2 Challenge Problem Design and Test Specification.* MITRE Technical Report, MTR140412.

Burns, K. (2014c). HELP: Formalizing frames in a story of sensemaking. Submitted to the *12th International Conference on Naturalistic Decision Making*, McLean, VA.

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

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

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

Burns, K., Fine, M., Bonaceto, C., & Beltz, B. (2014). *Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking (ICArUS): Phase 1 Challenge Problem Walkthrough*. MITRE Technical Report, MTR140413.

Burns, K., Greenwald, H., & Fine, M. (2014). *Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking (ICArUS): Phase 1 Challenge Problem Design and Test Specification*. MITRE Technical Report, MTR140410.

Carver, C., & White, T. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology, 67*, 319-333.

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.

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

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.

MITRE (2014). *Phase 2 Test and Evaluation Results*. Briefing presented at ICArUS Technical Exchange Meeting (TEM), June 3, 2014.

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

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

MITRE (2012b). *Results of the Phase 1 Experiment*. Briefing presented at ICArUS Technical Exchange Meeting (TEM), December 12, 2012.

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

Stech, F. (2014). Evolution of GEOINT doctrine and methods. Submitted to *Geospatial Intelligence Review*.

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