MTR140415
MITRE TECHNICAL REPORT

MITRE

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

A Computational Basis for ICArUS Challenge Problem Design

Kevin Burns

November, 2014

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

Approved for Public Release; Distribution Unlimited 14-3934

McLean, VA

Abstract

The IARPA (Intelligence Advanced Research Projects Activity) program ICArUS (Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking) requires "sensemaking" challenge problems as described in the BAA (Broad Agency Announcement, 2010). These problems are needed for conducting experiments with human participants, and for assessing neural-computational models of human performance in Test and Evaluation (T&E). Previously, researchers have described the cognitive challenges of sensemaking only informally using conceptual notions like "framing" and "re-framing", which are not sufficient to support T&E in accordance with the BAA. To overcome this limitation, a Bayesian-computational model of sensemaking was developed by dissecting a prototypical example of intelligence analysis, and by defining eight discrete steps in a processing cycle dubbed the Octaloop. This document details the Octaloop model and describes how it was used as a computational basis for design of ICArUS Phase 1 and Phase 2 challenge problems. Other uses of the Octaloop beyond ICArUS experiments are also identified. These uses include structured analytic techniques, training of critical thinking skills, and automated tools for improving the effectiveness of intelligence analysis.

Table of Contents

1 Introduction 3
1.1 A Conceptual Theory of Sensemaking 3
1.2 A Computational Approach to Sensemaking 5
2 Motivation 6
2.1 A Cycle Called the Octaloop 6
2.2 A Story of Sensemaking 6
3 Definitions 8
3.1 Summary of Key Terms 8
3.2 Glossary of All Terms 9
4 Formulation 12
4.1 A Cycle of Sensemaking 12
4.2 Eight Steps of the Octaloop 12
5 Demonstration 15
5.1 Suspecting "The Bad Guys" 15
5.2 Reviewing their Tactics 17
5.3 Abducting a Reason 18
5.4 Collecting More Data 19
5.5 Concluding "It's Students" 21
6 Application 23
6.1 The Importance of Likelihoods 23
6.2 Hypotheses and Evidence 24
6.3 The Nature of Re-Framing 24
6.4 Three Functions of Sensemaking 25
6.5 Key Insights on Biases 26
6.5.1 An Insight in Hindsight 26
6.5.2 The Benefits of Biases 27
6.5.3 Confirming Conservatism 27
6.5.4 Substitution of a Structured Technique 29
7 Transition 30
7.1 Technique to HELP Perform Bayesian Reasoning 30
7.1.1 Hypotheses 30
7.1.2 Evidence 31

7.1.3 Likelihoods 32
7.1.4 Posteriors 32
7.1.5 HELP 33
7.2 Training of Critical Thinking with Bayesian HELP 34
7.3 Tools for Cognitive Support to Analysts 35
7.4 Leveraging the Bayesian Research Community 37
8 References 39

1 Introduction

This document develops a computational model of sensemaking, dubbed the *Octaloop*, used to support the Intelligence Advanced Research Projects Activity (IARPA) program ICArUS: Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking. Further background on MITRE's role in Test and Evaluation for ICArUS is provided in a summary document (Burns, Fine, Bonaceto, & Oertel, 2014) titled *ICArUS: Overview of Test and Evaluation Materials*, available at http://www.mitre.org/publications.

The Octaloop model presented herein was developed as a formal basis for designing ICArUS challenge problems (Burns, Greenwald, & Fine, 2014; Burns, 2014), including methods for assessing the performance of human subjects and neural models on these challenge problems. Details of how the Octaloop was applied to the design of ICArUS challenge problems appear in Section 6. The Octaloop also offers opportunities for applications to real-world analytic techniques, training, and tools, as outlined in Section 7.

1.1 A Conceptual Theory of Sensemaking

ICArUS challenge problems are designed to pose cognitive challenges of spatial-temporal sensemaking, including six core processes described in the BAA (2010) as follows:

> **Learn Frames**: Construct mental models from the data; i.e., spatial context frames; event sequence frames (scripts); semantic relational frames.
>
> **Recognize Patterns / Select a Frame**: Based on current data, select the appropriate frame(s) from memory.
>
> **Assess the Frame**: Evaluate the quality of fit between data and frame.
>
> **Generate Hypotheses**: Use the current frame to generate hypotheses regarding missing data (either confirming or disconfirming) and to predict the future evolution of the data.
>
> **Acquire Additional Data**: Search for new data to test and complete the frame; assess value and uncertainty of data and data sources; decide whether to continue to exploit current data or to explore new sources.
>
> **Reframe**: Detect anomalies, coincidences, inconsistencies, and ambiguities in the data. Accept, modify, or reject frame as needed.

These core processes are based on a conceptual framework by Klein, et al. (2007), known as the *"data-frame theory"* of sensemaking, which addresses the human *"faculty to understand" ... "referred to as judgment, apprehension, apperception, and other processes"*.

The data-frame theory defines sensemaking as *"the deliberate attempt to understand events"* (pg. 114) and includes the following mental representations and processes (pg. 115):

> "*The initial account people generate to explain events.*
>
> *The elaboration of that account.*
>
> *The questioning of that account in response to inconsistent data.*
>
> *Fixation on the initial account.*
>
> *Discovering inadequacies in the initial account.*
>
> *Comparison of alternative accounts.*
>
> *Reframing the initial account and replacing it with another.*
>
> *The deliberate construction of an account when none is automatically recognized."*

As such, the theory includes three classes of mental representation: ***data***, ***frames***, and the ***accounts*** that result from mental processes of ***generating***, ***elaboration***, ***questioning***, ***fixation***, ***discovering***, ***comparison***, ***reframing***, and ***construction***. Klein, et al. (2007) offer more details in a section titled *"The Data-Frame Theory of Sensemaking"* (pg. 118), which states: *"The data-frame theory postulates that elements are explained when they are fitted into a structure that links them to other elements. We use the term* ***frame*** *to denote an explanatory structure that defines entities by describing their relationship to other entities"*. The same section goes on to say that: *"A frame can take the form of a story, ... map, ... script, ... plan, ... [or other] structure for accounting for the data and guiding the search for more data"* (pg. 118).

As described above and throughout the chapter by Klein, et al. (2007), the notions of ***data***, ***frames***, and ***accounts*** are not specified with the rigor needed to formally measure and model sensemaking. For example, it is not clear if ***entities*** and ***elements*** refer to ***data***, and/or ***frames***, and/or ***accounts*** – or how any of these things can be measured or modeled. Even the central notion of a ***frame*** itself is not precisely specified, as the theory says a frame might take the form of a story, map, script, plan, or any other explanatory knowledge structure.

In short, the data-frame theory does not offer a computational specification of the mental representations and processes involved in sensemaking. This limitation is important because the objective of ICArUS is to develop neural-computational models of sensemaking – and because the assessment of such models in Test and Evaluation (T&E) requires a computational framework for quantification. More specifically, the BAA (2010) requires that T&E assess human performance and neuroscience models in Comparative Performance Assessment (CPA), using a numerical percentage to measure how well a neural model matches human judgment and decision-making. Additionally, the BAA (2010) requires that T&E compute normative (Bayesian) solutions, as benchmarks for measuring cognitive biases of human subjects and neural models in Cognitive Fidelity Assessment (CFA).

1.2 A Computational Approach to Sensemaking

In light of the BAA requirements for CPA and CFA, MITRE developed a Bayesian model of sensemaking that entails eight steps called the Octaloop. To support CFA, the Octaloop specifies Bayesian-computational processes as a basis for assessing the biases that arise from heuristic reasoning under uncertainty. To support CPA, the Octaloop enables quantitative comparisons of sensemaking performance between human subjects and neural models.

Here it is important to note that a Bayesian approach is not contrary to the data-frame theory cited in the BAA. Rather, the Bayesian Octaloop serves to formalize notions such as *data*, *frames*, and the *accounts* produced by *generating*, *elaboration*, *questioning*, *fixation*, *discovering*, *comparison*, *reframing*, and *construction*. Also, the Bayesian Octaloop serves to formalize core sensemaking processes described in the BAA and listed in Section 1.1, which are based on the data-frame theory.

The remainder of this document is organized as follows: Section 2 describes a prototypical situation (taken from Klein, et al., 2007) in which an intelligence analyst was engaged in sensemaking. Section 3 defines key terms, such as *sensemaking*, *frames*, *framing*, and *re-framing*, as formalized in a Bayesian framework of *hypotheses*, *confidence*, *evidence*, and *likelihoods*. Section 4 details the mathematical formulation by which sensemaking is modeled in a cycle of eight steps dubbed the Octaloop. Section 5 demonstrates how the Octaloop applies to the situation in Section 2, by dissecting five consecutive cycles of sensemaking in the story. Section 6 discusses how the Octaloop has been applied in design of challenge problems that satisfy all requirements of the ICArUS BAA. Finally, Section 7 explores how the Octaloop can be applied beyond ICArUS challenge problems, to advance the techniques, training, and tools of real-world intelligence analysis.

2 Motivation

2.1 A Cycle Called the Octaloop

Below is a story of sensemaking in the context of intelligence analysis, borrowed from Klein, et al. (2007). Numbers that appear in brackets [ ] refer to steps 1-8 of the Octaloop model detailed in Section 4. The names of these eight steps are listed below for reference purposes:

[1] Isolating Evidence
[2] Generating Hypotheses
[3] Estimating Likelihoods
[4] Aggregating Confidence
[5] Prognosticating Consequence
[6] Evaluating Consequence
[7] Anticipating Evidence
[8] Discriminating Evidence

These eight steps constitute a cycle of sensemaking, which is typically repeated as additional evidence is accumulated, evaluated, and anticipated. In the narrative below, there are five cycles of sensemaking, each appearing in a separate paragraph. Typically one cycle ends and another cycle begins at steps [8] and [1] of the Octaloop.

2.2 A Story of Sensemaking

Major A. S. discussed an incident that occurred soon after 9/11 in which he was able to determine the nature of overflight activity around nuclear power plants and weapons facilities. This incident occurred while he was an analyst. He noticed [8] that there had been increased reports in counterintelligence outlets of overflight incidents around nuclear power plants and weapons facilities. At that time, all nuclear power plants and weapons facilities were "temporary restricted flight" zones. So this meant [1] there were suddenly a number of reports of small, low-flying planes around these facilities. At face value it appeared [2],[3],[4],[7] that this constituted a terrorist threat—that "bad guys" had suddenly increased their surveillance activities. There had not been any reports of this activity prior to 9/11 (but there had been no temporary flight restrictions before 9/11 either).

Major A. S. obtained [8] access to the Al Qaeda tactics manual, which instructed [1] Al Qaeda members not to bring attention to themselves. This piece of information helped him to begin to form [2] the hypothesis that these incidents were bogus—"It was a gut feeling, it just didn't sit right. If I was a terrorist I wouldn't be doing this." He recalled thinking [3],[4] to himself, "If I was trying to do surveillance how would I do it?" From the Al Qaeda manual, he knew they wouldn't break the rules, which to him meant that

they wouldn't break any of the flight rules. He asked himself, "If I'm a terrorist doing surveillance on a potential target, how do I act?" He couldn't put together a sensible story that had a terrorist doing anything as blatant as overflights in an air traffic restricted area.

He thought [2],[3],[4] about who might do that, and kept coming back to the overflights as some sort of mistake or blunder. That suggested student pilots to him because "basically, they are idiots." He was an experienced pilot. He knew that during training, it was absolutely standard for pilots to be instructed that if they got lost, the first thing they should look for were nuclear power plants. He told us that "an entire generation of pilots" had been given this specific instruction when learning to fly. Because they are so easily sighted, and are easily recognized landmarks, nuclear power plants are very useful for getting one's bearings. He also knew that during pilot training the visual flight rules would instruct students to fly east to west and low—about 1,500 feet. Basically students would [7] fly low patterns, from east to west, from airport to airport.

It took Major A. S. about 3 weeks to do his assessment. He found [8] all relevant message traffic by searching databases for about 3 days. He picked [1] the three geographic areas with the highest number of reports and focused on those. He developed overlays to show where airports were located and the different flight routes between them. In all three cases, the "temporary restricted flight" zones (and the nuclear power plants) happened to fall along a vector with an airport on either end. This added [2],[3],[4] support to his hypothesis that the overflights were student pilots, lost and using the nuclear power plants to reorient, just as they had been told to do. He also checked [7] to see if any of the pilots of the flights that had been cited over nuclear plants or weapons facilities were interviewed by the FBI.

In the message traffic, he discovered [8],[1] that about 10% to 15% of these pilots had been detained, but none had panned out as being "nefarious pilots." With this information, Major A. S. settled [2],[3],[4] on an answer to his question about who would break the rules: student pilots. The students were probably following visual flight rules, not any sort of flight plan. That is, they were flying by looking out the window and navigating.

3 Definitions

Section 4 will detail steps of the Octaloop, and Section 5 will demonstrate how it applies to the story of Section 2. But first, Section 3 defines key terms used to characterize sensemaking from a computational perspective.

3.1 Summary of Key Terms

Sensemaking is defined as follows, where words in ***bold italics*** appear alphabetically in the glossary of Section 3.2:

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

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

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

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

With that aim, the notion of a frame is formalized here as follows:

> ***Frames*** are knowledge structures comprising ***hypotheses***, ***evidence***, ***confidences*** in ***hypotheses***, and ***likelihoods*** of ***evidence***.

Notice that this definition of a frame goes beyond that of the data-frame theory, because here a frame always represents data (i.e., ***evidence***) as well as other knowledge and beliefs (i.e., ***hypotheses***, ***likelihoods***, and ***confidences***) needed to make sense of data. The reason is that likelihoods are needed for computing confidences in hypotheses, and likelihoods always refer to data (evidence) – simply because a likelihood is the probability of evidence given a hypothesis.

The data-frame theory instead suggests a frame is everything except data that is needed to make sense of data. Besides excluding data from the frame, that definition says only what a frame is not and fails to specify what a frame is. The more formal definition above implies a frame cannot exclude data, if the frame is to be useful for making sense of data. The above definition also specifies exactly what else besides ***evidence*** (data) is needed in order to make sense of a situation, namely ***hypotheses***, ***likelihoods***, and ***confidences***. Only when all these components of frames are made explicit, as in the formal definition above, is it possible to compute how frames might be "learned" and "assessed" and "re-framed" per the BAA's core sensemaking processes listed in Section 1.1. In particular, the notions of ***framing*** and ***re-framing*** are formalized here as follows:

> ***Framing*** is the formation of ***confidences*** across ***hypotheses***, based on ***evidence*** and the ***likelihoods*** of that ***evidence*** being caused by various ***hypotheses***.
>
> ***Re-framing*** is a revision of ***hypotheses*** and/or ***confidences*** across ***hypotheses***, based on evidence and the ***likelihoods*** of that ***evidence*** being caused by various ***hypotheses***.

According to these definitions, there are two differences between framing and re-framing. First, framing is an initial formation of confidences across hypotheses, whereas re-framing is the subsequent revision of a frame formed earlier. Second, in re-framing some components of the frame may not be re-formed. In particular, re-framing may involve a revision of ***confidences*** across a set of hypotheses without changing the categorical set of ***hypotheses*** itself. Later Section 5 will identify this type of re-framing across fixed hypotheses in the sensemaking story of Section 2, as well as a type of re-framing in which new hypotheses are generated.

In all the above definitions, a distinction between ***hypotheses*** and ***evidence*** is especially important – because it reflects the causal structure (Pearl, 2000) that underlies sensemaking in both ***explanation*** and ***prediction***. This causal structure is: hypotheses → evidence; where hypotheses are possible causes of evidential effects (i.e., causes → effects), and inferencing can proceed either forward (in the direction of the arrow) or backward. In backward inferencing, a sensemaker is explaining evidence. In forward inferencing, a sensemaker is predicting evidence. The distinction between ***confidence*** and ***likelihood*** then parallels the distinction between ***hypotheses*** and ***evidence***. That is, confidence refers to the probability of a hypothesis given some evidence, whereas likelihood refers to the probability of some evidence given a hypothesis.

3.2 Glossary of All Terms

Below is a more comprehensive set of definitions, in alphabetical order, including all of the terms that appear in ***bold italics*** above:

> ***Bayesian*** refers to the use of Bayes Rule for updating beliefs in ***hypotheses*** given ***evidence***. Bayes Rule is mathematical specification for how ***prior probabilities*** of ***hypotheses*** and conditional ***likelihoods*** of ***evidence*** are combined to compute ***posterior probabilities*** of ***hypotheses***. ***Bayesian*** also refers to the optimal computation of expected

utility in decision-making situations, as the product of ***probability*** and ***utility*** summed across all possible outcomes of an action that may be chosen.

Confidence is a measure of belief in the truth of a ***hypothesis***, given some ***evidence*** (or ***prior*** to some ***evidence***). ***Confidence*** is quantified by ***probability***.

Evidence is factual information (i.e., data) about an uncertain situation, obtained by direct observation or some communication (e.g., from a source of intelligence).

Explanations are backward inferences to compute ***confidences*** in ***hypotheses***, given ***evidence***.

Frames are knowledge structures comprising ***hypotheses***, ***evidence***, ***confidences*** in ***hypotheses***, and ***likelihoods*** of ***evidence***.

Framing is the formation of ***confidences*** across ***hypotheses***, based on ***evidence*** and the ***likelihoods*** of that ***evidence*** being caused by various ***hypotheses***.

Hypotheses are possible ***explanations*** of ***evidence***, involving causal reasons for ***evidence*** as reflected by ***likelihoods***.

Inferencing is the assignment of ***confidences*** to ***hypotheses*** (backward inferencing) or the assessment of ***likelihoods*** for ***evidence*** (forward inferencing).

Likelihood is a measure of belief in the occurrence of some ***evidence***, given a ***hypothesis***. ***Likelihood*** is quantified by ***probability***.

Posterior refers to the result of a ***Bayesian*** update, in which ***prior probabilities*** (of ***hypotheses***) are combined with conditional ***likelihoods*** (of ***evidence*** given ***hypotheses***) in order to compute ***posterior probabilities*** (of ***hypotheses*** given ***evidence***). This is a backward ***inference***, resulting in the ***explanation*** of a situation.

Predictions are forward inferences to compute ***likelihoods*** of ***evidence***, given ***hypotheses***.

Prior refers to the ***probability*** of a ***hypothesis*** before obtaining some ***evidence***.

Probability is a mathematical measure of belief in the truth of a ***hypothesis*** or occurrence of some ***evidence***, quantified by a number in the range 0-1.

Re-framing is a revision of ***hypotheses*** and/or ***confidences*** across ***hypotheses***, based on evidence and the ***likelihoods*** of that ***evidence*** being caused by various ***hypotheses***.

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

Utility is a mathematical measure of value or consequence for the outcome of an action that may or may not be chosen among a set of possible actions.

4 Formulation

4.1 A Cycle of Sensemaking

Based on the formal definitions of Section 3, and the anecdotal motivation in Section 2, a cycle of sensemaking can be characterized in terms of *what is computed* at each step of the cycle. Previous research (Burns, in press; 2012; 2011; 2010; 2007; 2006; 2005) has shown that a Bayesian approach (Edwards, 1961; 1954; Fischhoff & Beyth-Marom, 1983) provides a useful framework for analyzing the process of sensemaking (and its product called "situation awareness", see Klein, et al., 2007, pp. 119-120) from a computational perspective, including various heuristics and biased deviations relative to Bayesian norms.

Using the Bayesian approach, a cycle of sensemaking is formally expressed in the following eight steps dubbed the *Octaloop*:

4.2 Eight Steps of the Octaloop

[1] Isolating Evidence: After [8] Discriminating Evidence from the previous cycle, a sensemaker M decides which evidence e among all perceived evidence {e} is to be used in the current cycle of sensemaking. Some evidence in {e} is necessarily ignored, at least temporarily. The selection of e is a matter of attention, and is affected by M's current understanding of the situation after the previous cycle of sensemaking.

[2] Generating Hypotheses: After (or perhaps while) isolating evidence e, M creates or maintains a set of hypotheses $\{H_i\}$ that serve as possible explanations of e. M also represents a prior confidence $P(H_i)$ in each hypothesis, before assessing how the evidence e affects confidence across the set of hypotheses $\{H_i\}$. The values of $\{P(H_i)\}$ satisfy Σ_i $P(H_i) = 1$, such that each $P(H_i)$ represents a relative confidence in the truth of hypothesis H_i within the set of hypotheses $\{H_i\}$.

[3] Estimating Likelihoods: Along with each hypothesis H_i and prior $P(H_i)$, M also represents each likelihood $P(e|H_i)$ that the evidence e would be observed if the hypothesis H_i was true. Initial estimates for these likelihoods would arise at step [2] Generating Hypotheses, because likelihoods of the form $P(e|H_i)$ are what govern the generation of hypotheses H_i upon observing evidence e. However, further cogitations are often involved in refining the estimates of likelihoods, after a set $\{H_i\}$ of hypotheses has been established.

[4] Aggregating Confidence: Using the priors $P(H_i)$ and likelihoods $P(e|H_i)$, for all H_i in the set $\{H_i\}$, M updates his prior beliefs to obtain posterior beliefs $P(H_i|e)$. These posteriors are computed as normalized products of priors and likelihoods, per Bayes Rule

as follows: $P(H_i|e) = P(H_i) * P(e|H_i) / P(e)$, where the denominator is a normalizing factor computed as: $P(e) = \Sigma_i P(H_i) * P(e|H_i)$, and the sum is taken over all hypotheses in the set $\{H_i\}$. This normalizing factor ensures that the posterior probabilities (like the prior probabilities) sum to 1, i.e., $\Sigma_i P(H_i|e) = 1$.

[5] Prognosticating Consequence: Sometimes the sensemaker M can recommend or implement actions that affect the operational situation (e.g., to defend against threats). Any outcome of an action will depend on the actual state of the situation, which is uncertain and modeled by M's posterior beliefs $\{P(H_i|e)\}$. Thus using these posteriors to model potential states of the actual world, and given a set of possible actions $\{a_j\}$ that might be taken, M predicts the chances $P(o_k|a_j)$ of various outcomes (o_k) for each possible action (a_j). M also assigns a value or consequence to each outcome, modeled mathematically by utility $U(o_k|a_j)$. Finally, the Bayesian decision is to choose the action a_j that maximizes expected utility X_j, computed for each a_j as follows: $X_j = \Sigma_k P(o_k|a_j) * U(o_k|a_j)$.

[6] Evaluating Consequence: After taking an action that affects the situation, M observes the consequence and has a reaction. For example, the reaction to an outcome o_k that was expected at probability $P(o_k)$ would be quantified by Shannon's (1949) information-theoretic measure of surprisal, $-\log P(o_k)$. In effect an outcome o_k represents further information about the situation, and M's reaction to o_k may affect subsequent steps and cycles of sensemaking.

[7] Anticipating Evidence: Based on the current hypotheses $\{H_i\}$ and posteriors $\{P(H_i|e)\}$, along with likelihoods of the form $P(f|H_i,e)$ estimated for future evidence f that may be observed, M forms expectations (i.e., predictions) about each f in a set of possible evidence $\{f\}$ as follows: $P(f) = \Sigma_i P(H_i|e) * P(f|H_i,e)$, where the sum is taken over all hypotheses in the set $\{H_i\}$. Sometimes M has control over which evidence will be obtained next, and he makes a choice among options in a process known as "foraging" for information (Pirolli, 2007). The optimal choice would maximize expected utility much like step [5], which was a decision to take action that directly impacts the operational situation. But step [7] differs in that the actions a_j are options for collecting information, and the outcomes o_k are resulting gains in information. Thus probabilities $P(o_k|a_j)$ represent the chances of various outcomes for each action (i.e., to get information), and utilities $U(o_k|a_j)$ represent the associated values of those outcomes (i.e., gains in information). For example, the information gain $U(o_k|a_j)$ can be computed (Burns, 2011) in terms of entropy divergence (Kullback & Leibler, 1951) across a set of hypotheses $\{H_i\}$, measured from before an action (a_j) to after an outcome (o_k) of that action.

[8] Discriminating Evidence: After anticipating evidence, M obtains actual evidence (via active collection or passive observation) and represents this new evidence $\{e'\}$ for the next cycle of sensemaking. The set $\{e'\}$ is an interpretation of raw data, and may be uncertain, so in fact it might be considered a set of hypotheses $\{E'\}$. If so, then a secondary cycle of sensemaking would be embedded within step [8], in order to establish the set of evidence $\{e'\}$ that will be considered factual and carried forward to step [1].

Clearly some steps of this cycle are extremely complex. For example, step [8] may involve a secondary cycle of sensemaking within the primary cycle, in order to obtain {e'} from {E'}. Similarly, steps [7] and [1] may involve meta-sensemaking, where M would model his own sensemaking processes in order to optimize the collection of evidence at step [7] and selection of evidence at step [1]. By necessity, these and other complexities of real-world sensemaking are greatly simplified in ICArUS challenge problems – as discussed further in Section 6.

As outlined above, the Octaloop is a normative (Bayesian) model of the sensemaking cycle. However this same framework can be used as a descriptive model of cognitive (human) processing, simply by treating human heuristics as naturally bounded approximations to Bayesian computations. In fact the Bayesian framework is ideal for measuring and modeling psychological biases, precisely because it is a Bayesian framework and hence suitable for computing normative benchmarks as needed to measure subjective biases.

5 Demonstration

Section 2 described an anecdotal story, Section 3 defined more formal concepts, and Section 4 detailed a computational model. Section 5 now maps this model back to the story in order to demonstrate how sensemaking can be analyzed from a computational perspective. The story contains five cycles of sensemaking, each addressed in a separate subsection below.

5.1 Suspecting "The Bad Guys"

> *Major A. S. discussed an incident that occurred soon after 9/11 in which he was able to determine the nature of overflight activity around nuclear power plants and weapons facilities. This incident occurred while he was an analyst. He noticed [8] that there had been increased reports in counterintelligence outlets of overflight incidents around nuclear power plants and weapons facilities. At that time, all nuclear power plants and weapons facilities were "temporary restricted flight" zones. So this meant [1] there were suddenly a number of reports of small, low-flying planes around these facilities. At face value it appeared [2],[3],[4],[7] that this constituted a terrorist threat—that "bad guys" had suddenly increased their surveillance activities. There had not been any reports of this activity prior to 9/11 (but there had been no temporary flight restrictions before 9/11 either).*

The first cycle begins as the sensemaker (hereafter denoted M) is [8] Discriminating Evidence and [1] Isolating Evidence, by identifying a body of evidence {e} from counterintelligence and attending to an item of evidence e, denoted here as ***s = sudden increase (after 9/11) in reported flight zone violations.*** M thought this evidence constituted a terrorist threat, so he was [2] Generating Hypotheses $\{H_i\}$ about possible causes of the evidence s, and [3] Estimating Likelihoods of the form $P(s|H_i)$ for various H_i.

These likelihoods govern which hypotheses are recalled from long-term memory and represented in working memory as possible explanations of s. The story mentions a hypothesis ***A = Al Qaeda***, which suggests that evidence s was strongly associated with A, i.e., A was a likely cause of s. Besides this likelihood P(s|A), M would also be representing a prior probability P(A) that reflects his preconceived confidence in A, i.e., in the absence of evidence s.

Although the story does not mention it, the sensemaker M would have generated other hypotheses besides A. At the very least M must have generated the hypothesis ***~A = Not Al Qaeda***, because he was clearly not certain that the evidence s was caused by A. So in this first cycle of sensemaking, M would be representing at least two hypotheses {A, ~A} in his working memory. He would also be representing two priors P(A) and P(~A), and two likelihoods P(s|A) and P(s|~A).

Unfortunately this story, like most stories, does not provide numerical values for any of these probabilities. And if asked, the sensemaker M might even deny that he represented such values

in his mind. But actually M must mentally represent the probabilities at least implicitly, simply because he is not equally confident in A and ~A. For example, P(s|A) would be represented as the associative strength between s and A, where A is a possible cause of s. This strength differs from the associative strength between s and ~A, which is smaller because M knows a reason why A would cause s but does not know a reason why ~A would cause s. The point here is that the magnitudes of these associative strengths must be represented at least implicitly in the mind of a sensemaker, even if they are not reported explicitly as probabilities.

Here, for purposes of quantification, we can assign numbers that are consistent with the narrative of the story. For priors, we might assume P(A) = P(~A) = 0.50 if M had no prior beliefs about the probabilities of A or ~A. However, the events of the story took place soon after the 9/11 attacks, and the narrative suggests that M may have thought P(A) > P(~A). Thus for purposes of quantification, the analysis here will assume P(A) = 0.80 and P(~A) = 0.20. Note that P(A) + P(~A) = 1, because A and ~A are mutually exclusive and exhaustive hypotheses.

Also consistent with the story, we might assume P(s|A) = 0.90 and P(s|~A) = 0.50 for the likelihood of observing the evidence s if A or ~A were true, respectively. But notice that, unlike the priors, these likelihoods need not (and usually will not) sum to 1. Instead P(s|A) + P(~s|A) = 1, because if A is true then either s or ~s would occur. Thus the assumed value P(s|A) = 0.90 and implied value P(~s|A) = 1 - 0.90 = 0.10 together mean that M thinks Al Qaeda is much more likely to cause s than ~s, because M can think of a "reason" why A would cause s rather than ~s. Similarly, P(s|~A) + P(~s|~A) = 1, because if ~A is true then either s or ~s would occur. Here the assumed value P(s|~A) = 0.5 means that s would be a non-causal or "random" (i.e., for no apparent reason) effect if ~A was true, such that P(s|~A) = P(~s|~A) = 0.50.

After [3] Estimating Likelihoods as described above, the cycle continues as M engages in [4] Aggregating Confidence. According to the story M is led to believe that the evidence s was most likely caused by the "bad guys" (A). This belief can be quantified as a normalized product of priors and likelihoods, computed for each hypothesis (A and ~A) via Bayes Rule as follows:

$$P(A|s) = P(A) * P(s|A) / P(s)$$

$$P(\sim A|s) = P(\sim A) * P(s|\sim A) / P(s)$$

where P(s) is a normalizing factor appearing in the denominators, computed from the sum of numerators as follows:

$$P(s) = P(A) * P(s|A) + P(\sim A) * P(s|\sim A).$$

Using the numbers noted above, these equations produce posterior probabilities of P(A|s) = 0.88 and P(~A|s) = 0.12. In words, M would be thinking the most probable explanation of s is Al Qaeda's surveillance activities. Then based on this belief, M would be [7] Anticipating Evidence as he prepares to enter the next cycle of sensemaking. This anticipation would affect whether and

where he would seek further evidence. It would also affect how he directs his attention in [8] Discriminating Evidence after evidence is received.

This concludes the first cycle of sensemaking, which touched on all steps of the Octaloop except [5] Prognosticating Consequence and [6] Evaluating Consequence. These two steps refer to operational actions and outcomes, rather than analytical judgments and beliefs. The story is about the sensemaking of an intelligence analyst M, and does not involve any operational actions by M or others that he may advise to affect the operational situation. But more generally sensemaking may include both analytical and operational components, and that is why steps [5] and [6] are included in the Octaloop.

Similar to the choices and outcomes of operational actions in steps [5] and [6], steps [7] and [8] address choices and outcomes of analytical actions to obtain and exploit further information. The main difference lies in the modeling of utility, which for [5] and [6] concerns the value to operations, but for [7] and [8] concerns the value of information. According to the story, M's belief that A was probably true led him to seek further information about A from the Al Qaeda manual. Per the story:

5.2 Reviewing their Tactics

> *Major A. S. obtained [8] access to the Al Qaeda tactics manual, which instructed [1] Al Qaeda members not to bring attention to themselves. This piece of information helped him to begin to form [2] the hypothesis that these incidents were bogus—"It was a gut feeling, it just didn't sit right. If I was a terrorist I wouldn't be doing this." He recalled thinking [3],[4] to himself, "If I was trying to do surveillance how would I do it?" From the Al Qaeda manual, he knew they wouldn't break the rules, which to him meant that they wouldn't break any of the flight rules. He asked himself, "If I'm a terrorist doing surveillance on a potential target, how do I act?" He couldn't put together a sensible story that had a terrorist doing anything as blatant as overflights in an air traffic restricted area.*

Here a second cycle of sensemaking begins as M is [8] Discriminating Evidence and [1] Isolating Evidence. Although the story does not say so explicitly, these steps occur after an implicit step [7] Anticipating Evidence in the previous cycle of sensemaking, because clearly M had some expectation about what he would find in the Al Qaeda manual. Apparently he expected the manual would say something that would shed light on the likelihood P(s|A). But we cannot tell from the story if he expected to learn something that would increase or decrease his estimate of P(s|A).

Notice that the evidence (Al Qaeda manual) in this cycle is known to be caused by A rather than ~A. Therefore this evidence actually represents new knowledge about a likelihood P(s|A) that was estimated in the previous cycle of sensemaking. Based on his review, M now thinks that P(s|A) is very small, because the manual instructs Al Qaeda members not to bring attention to themselves. For example, perhaps after reading the manual M thought P(s|A) = 0.01. In effect, M

realized that his previous estimate of P(s|A) = 0.90 was wrong, because he learned of a very good reason (from the Al Qaeda manual) for why A would not cause s (and instead would cause ~s). So M repeats the previous cycle of sensemaking, but now using P(s|A) = 0.01 instead of P(s|A) = 0.90. The Al Qaeda manual says nothing about other groups (~A), so P(s|~A) remains = 0.50.

Using the revised likelihoods, along with the original priors P(A) = 0.80 and P(~A) = 0.20, the Bayesian equations produce posteriors as follows: P(A|s) = 0.07 and P(~A|s) = 0.93. In words, the sensemaker's beliefs have undergone a reversal, from A being very probable to ~A being very probable, based on a change in the likelihood P(s|A). This is an example of ***re-framing*** in which additional evidence (from the Al Qaeda manual) changes M's estimate of a likelihood, and this in turn changes M's beliefs about the most probable explanation of previous evidence s (from counterintelligence) across a set of hypotheses {A, ~A}.

As a result, the story says that M *"began to form the hypothesis that these incidents were bogus"*. But notice that this is not really a new hypothesis, because the hypothesis ~A had been generated earlier along with the hypothesis A. Instead at this point M merely began to think that ~A was more probable than A, according to the Bayesian calculations outlined above, which correspond to steps [3] Estimating Likelihoods and [4] Aggregating Confidence of the Octaloop.

Also at this point the story says that M began to wonder who, besides a terrorist, would possibly break the rules and hence cause the observed evidence s. Eventually M generated a novel hypothesis in answer to this question, but it was not until the next cycle of sensemaking. What is interesting here, in the present cycle, is that M felt compelled to wonder about the hypothesis ~A and eventually generate a new hypothesis (regarding student pilots). It appears that M's motivation for doing so was twofold. First, he now thought ~A was the most probable hypothesis. Second, his likelihoods for this most probable hypothesis (~A) were P(s|~A) = 0.50 and P(~s|~A) = 0.50, so he had no causal basis or reason by which he could explain the evidence s. In other words, M was pretty sure he knew who was <u>not</u> responsible for the overflight activity, but he still did not have a clue as to who <u>was</u> responsible. And apparently he felt a strong need to establish who <u>was</u> probably responsible, rather than who was <u>not</u> probably responsible.

5.3 Abducting a Reason

> *He thought [2],[3],[4] about who might do that, and kept coming back to the overflights as some sort of mistake or blunder. That suggested student pilots to him because "basically, they are idiots." He was an experienced pilot. He knew that during training, it was absolutely standard for pilots to be instructed that if they got lost, the first thing they should look for were nuclear power plants. He told us that "an entire generation of pilots" had been given this specific instruction when learning to fly. Because they are so easily sighted, and are easily recognized landmarks, nuclear power plants are very useful for getting one's bearings. He also knew that during pilot training the visual flight rules would instruct students to fly east to west and low—about 1,500 feet. Basically students would [7] fly low patterns, from east to west, from airport to airport.*

Motivated by his desire to find a better (causal) explanation for the evidence s, M initiated this third cycle of sensemaking without the introduction of any new evidence from steps [8] and [1]. That is, M engaged in [2] Generating Hypotheses about who was responsible for s, after realizing that Al Qaeda (A) was probably not responsible.

The result is a new hypothesis ***S = Student pilots (and not Al Qaeda)***, based on high associative memory strength between S and s. This strength, in turn, reflects a reason (i.e., knowledge) for why S would cause s. That is, given M's expert knowledge as a pilot, he thinks P(s|S) is high because he knows why students would frequently fly over nuclear power plants. Numerically, we might assume P(s|S) = 0.90 because students have a reason for causing s, whereas P(s|~S) = 0.50 because non-students may or may not have a reason for causing s.

At this point M's set of hypotheses can be characterized as {A, S, ~S), where ***~S = Not student pilots (and not Al Qaeda)***. So here the ***re-framing*** is more comprehensive than we saw in the previous cycle, because here we have new hypotheses as well as new likelihoods associated with those hypotheses. For priors, we can assume P(A) = 0.80 as before, and then assume P(~A) = 0.20 is split equally between the two hypotheses that were not previously distinguished within ~A such that P(S) = P(~S) = 0.10. For likelihoods, we have P(s|A) = 0.01 from the previous cycle of sensemaking, and now from step [3] of the present cycle we have P(s|S) = 0.90 and P(s|~S) = 0.50. Finally after [3] Estimating Likelihoods in this fashion, Bayes Rule is once again used for [4] Aggregating Confidence.

The resulting posteriors are: P(A|s) = 0.05, P(S|s) = 0.61, and P(~S|s) = 0.34. In words, M thinks S is about ten times more probable than A, and he also thinks S is about twice as probable as ~S. As in an earlier cycle of sensemaking, where M thought to consult the Al Qaeda manual, his beliefs here lead to [7] Anticipating Evidence that can better distinguish the cause (A, S, or ~S) of evidence s. Also as in earlier cycles, the story does not say why M chose to collect the evidence about flight paths (analyzed in the next cycle of sensemaking), but presumably he expected this evidence would help in establishing relative confidence in S versus ~S.

5.4 Collecting More Data

> *It took Major A. S. about 3 weeks to do his assessment. He found [8] all relevant message traffic by searching databases for about 3 days. He picked [1] the three geographic areas with the highest number of reports and focused on those. He developed overlays to show where airports were located and the different flight routes between them. In all three cases, the "temporary restricted flight" zones (and the nuclear power plants) happened to fall along a vector with an airport on either end. This added [2],[3],[4] support to his hypothesis that the overflights were student pilots, lost and using the nuclear power plants to reorient, just as they had been told to do. He also checked [7] to see if any of the pilots of the flights that had been cited over nuclear plants or weapons facilities were interviewed by the FBI.*

In this fourth cycle of sensemaking, M begins by [8] Discriminating Evidence and [1] Isolating Evidence. He found that vectors through restricted zones had airports on either end, and the story says this evidence added support to his hypothesis (S). But actually the evidence first affected his estimates of likelihoods, which in turn affected his confidence in each hypothesis {A, S, ~S}. More specifically, M's finding that some vectors between airports passed directly over nuclear power plants led him to increase the likelihood P(s|S) and decrease the likelihood P(s|~S), relative to his earlier estimates for these same likelihoods. In that respect the ***re-framing*** here is much like that in the second cycle, where reading the Al Qaeda manual led M to decrease the likelihood of P(s|A).

For example, based on his geospatial analysis, perhaps M increased the likelihood of P(s|S) from 0.90 to 0.95, and decreased the likelihood of P(s|~S) from 0.50 to 0.10. The increase in P(s|S) reflects M's finding of airport vectors over nuclear plants, which increased his belief that S would cause s. The decrease came from his finding of other flight paths that would presumably be used by experienced pilots, which decreased his belief that ~S would cause s. Notice that the decrease in P(s|~S) is more drastic than the increase in P(s|S), because P(s|S) = 0.90 was already high, whereas previously P(s|~S) = 0.50.

Assuming the revised likelihoods are P(s|A) = 0.01, P(s|S) = 0.95, P(s|~S) = 0.10, and using the previous cycle's priors P(A) = 0.80, P(S) = 0.10, and P(~S) = 0.10, the Bayesian posteriors are computed are follows: P(A|s) = 0.07, P(S|s) = 0.84, and P(~S|s) = 0.09. In words, M now thinks that S is about ten times more probable than either A or ~S. Thus after developing the flight path overlays (and reviewing the Al Qaeda manual), M is even more certain that the most probable explanation for the overlflight activity is student pilots (who are not members of Al Qaeda).

As in previous cycles of sensemaking, these beliefs lead M to seek further information that will help establish relative confidence in competing hypotheses. Once again, the story does not say exactly why M decides to check the FBI records. Perhaps he thought the records might say whether violators were students or not (i.e., S versus ~S); or perhaps he thought the records would uncover any ties that pilots had to Al Qaeda (i.e., A versus ~A).

One interesting aspect of this story is that M chose to spend days/weeks on the overlay analysis before checking the FBI records. Assuming that M's primary concern was the threat of Al Qaeda activity, and therefore the probability of A versus ~A, it is curious that he first chose to perform the overlay analysis, which presumably would help distinguish S from ~S but not help distinguish A from ~A. So although the story does not say, it appears that M felt the probability of A was low enough, and he was more concerned with finding evidence to support his belief that S was the most likely cause of s. In other words, M's priority for further analysis was to establish what did cause s (which he suspected was S), rather than what did not cause s.

This behavior might be characterized as a *Confirmation Bias* (Nickerson, 1998), because M chooses to collect evidence that pertains to a more probable (and less consequential) hypothesis S, rather than collect evidence that pertains to a less probable (and more consequential) hypothesis A. But it is not clear whether the behavior was actually non-normative (sub-optimal) or not. And the answer to that question would require that many more parameters of the situation be identified and quantified – including each option available to the sensemaker for collecting

information, along with the expected costs and benefits (of what might be learned from further information) with respect to the analytical functions and operational missions that M's sensemaking might support.

5.5 Concluding "It's Students"

> *In the message traffic, he discovered [8],[1] that about 10% to 15% of these pilots had been detained, but none had panned out as being "nefarious pilots." With this information, Major A. S. settled [2],[3],[4] on an answer to his question about who would break the rules: student pilots. The students were probably following visual flight rules, not any sort of flight plan. That is, they were flying by looking out the window and navigating.*

In this fifth and final cycle of sensemaking, the result of [8] and [1] was a finding of: ***n = no nefarious pilots identified in the FBI interviews***. Here the evidence n is like s in that it could feasibly be caused by any of the candidate hypotheses: {A, S, ~S}. This differs from the evidence considered in the second cycle (Al Qaeda manual), which pertains only to A; and evidence considered in the fourth cycle (flight path analysis), which pertains only to ~A (i.e., S and ~S). Here in the fifth cycle M has new evidence that pertains to each hypothesis as he performs step [3] Estimating Likelihoods and step [4] Aggregating Confidence.

The new likelihoods that must be estimated are probabilities of evidence n, conditional on each hypothesis {A, S, ~S} but also conditional on the previous evidence s. Because n presumably comes from a different and diverse source of intelligence than the counterintelligence reports s, we can assume that n and s are independent. Thus the likelihoods of n are conditioned only on hypotheses, as follows: P(n|A), P(n|S), and P(n|~S). For example, based on the sample of pilots that had been interviewed, a finding of no nefarious pilots might suggest P(n|A) = 0. But because the sample is limited to 10-15% of pilots, and because interviews of pilots would not be 100% reliable in establishing ties to Al Qaeda, M might assume P(n|A) = 0.01 and P(~n|A) = 0.99. On the other hand, it appears the FBI data were uninformative with respect to the student status of pilots. So for students we have P(n|S) = 0.50 and P(~n|S) = 0.50, and also for non-students we have P(n|~S) = 0.50 and P(~n|~S) = 0.50.

Armed with these three likelihoods, P(n|A) = 0.01, P(n|S) = 0.50, and P(n|~S) = 0.5, Bayes Rule is used to update the posteriors computed in the previous cycle of sensemaking. Those posteriors become priors in the present cycle: P(A|s) = 0.07, P(S|s) = 0.84, and P(~S|s) = 0.09. Combining these priors with the likelihoods via Bayes Rule we obtain the following posteriors: P(A|n,s) = 0.001, P(S|n,s) = 0.90, and P(~S|n,s) = 0.10. In words, after five cycles of sensemaking the sensemaker M is now very sure the evidence (s and n) is <u>not</u> explained by Al Qaeda activity, P(A|n,s) = 0.001. He is also pretty sure that the evidence <u>is</u> explained by activities of student pilots following visual flight rules, P(S|n,s) = 0.90.

Notice that this fifth cycle of sensemaking involved a ***re-framing*** unlike the other two types we have seen. That is, there are no changes to any previously-estimated likelihoods, and there are no

newly-generated hypotheses. Instead there is new evidence (n) from an independent source and associated likelihoods P(n|A), P(n|S), and P(n|~S) across a fixed set of hypotheses {A, S, ~S). The result is a new set of posteriors that reflect the aggregation of original priors (from the first cycle of sensemaking) and likelihoods of all evidence (counterintelligence reports, Al Qaeda manual, flight path overlays, and FBI interviews).

6 Application

Section 5 demonstrated how real-world sensemaking could be analyzed from a computational perspective, using the Octaloop. Section 6 now discusses how the Octaloop has been used as a basis for designing ICArUS "challenge problems", which pose prototypical challenges of sensemaking within constraints imposed by the BAA.

The discussion here is intended to be general and hence does not address details of individual tasks or trials of the Phase 1 or Phase 2 challenge problems. For those details, readers are referred to the Phase 1 design document (Burns, Greenwald, & Fine, 2014) and the Phase 2 design document (Burns, 2014).

6.1 The Importance of Likelihoods

One of the most important insights from Section 5 is that ***likelihoods*** are central to sensemaking. This is noteworthy because likelihoods are not mentioned or modeled in the "data-frame" theory, or the core sensemaking processes identified in the BAA, yet they are critical components (along with ***hypotheses***, ***evidence***, and ***confidences***) of ***frames***. Any sensemaker in the real-world or laboratory must mentally represent likelihoods, at least implicitly. Unfortunately, likelihoods are usually not expressed explicitly or measured numerically, as we saw in the prototypical story of Section 2. And in that case it is impossible to model and measure sensemaking relative to normative (Bayesian) standards – as needed to gain a computational understanding of sensemaking, and as needed to assess performance and biases per the ICArUS BAA.

Basically ICArUS experiments must either provide likelihoods as inputs to human subjects (and neural models), or else measure the likelihoods that are being used by those human subjects (and neural models) as they make sense of evidence. As a practical matter, it is onerous for human subjects to report all likelihoods in each cycle of sensemaking. And even if human subjects would be willing to do so, it is not feasible for experimenters to measure "average" human performance (per the ICArUS BAA) when all subjects are using their own personal estimates of likelihoods.

Finally, even if it were feasible, each subject's personal likelihoods would be affected by the real-world knowledge that he or she brings to an experiment – much like the story in which a sensemaker estimated the likelihoods of flight zone violations based on his own experience as a pilot. The ICArUS BAA requires that challenge problems be designed to minimize the effects of rich and sophisticated knowledge representations (RASKRs) held by human subjects, because neural models being built and tested in the program will not possess the same expert knowledge.

For these reasons, likelihoods are given to humans and models as input to sensemaking in tasks of the ICArUS Phase 1 and Phase 2 challenge problems. In effect, subjects are provided with the results of step [3] Estimating Likelihoods, as if this step were being performed by a teammate or system rather than as part of their own sensemaking cycle.

6.2 Hypotheses and Evidence

By definition (per Section 3), likelihoods are probabilities of the form P(e|H). So if a subject is to be given all the likelihoods needed for sensemaking (per Section 6.1), then by necessity the subject must also be given the evidence (e) and hypotheses $\{H_i\}$ in each cycle of sensemaking. Thus like step [3] Estimating Likelihoods, subjects must be provided with the results of steps [8] Discriminating Evidence, [1] Isolating Evidence, and [2] Generating Hypotheses, rather than performing these steps themselves.

This may make ICArUS challenge problems feel somewhat unnatural to participants in experiments, as analysts are used to performing all steps themselves or in concert with human teammates rather than machine systems. Participants are also not used to reporting intermediate results, especially not in numerical fashion, which they must also do in ICArUS experiments. However these experimental controls are required to enable rigorous modeling and measuring of sensemaking per the ICArUS BAA.

Regarding steps [8] and [1], there is another reason for providing subjects with features of evidence (i.e., data) rather than requiring that they extract those features from raw sensory representations. The reason is that low-level visual perception and natural language processing are excluded from the scope of the ICArUS program. Like RASKRs discussed above, ICArUS models are not being developed to solve problems of vision or language. And because neural models will not have those capabilities, human subjects should not exploit their own capabilities for vision and language if comparisons between humans and models are to be meaningful.

6.3 The Nature of Re-Framing

In analyzing the story of sensemaking, Section 5 identified three different types of ***re-framing***. These three types can be characterized as *Abducting*, *Revising*, and *Updating*. *Abducting* occurred in the original framing that generated hypotheses in the set {A, ~A}, then again in later re-framing that generated new hypotheses in an expanded set {A, S, ~S}. *Revising* occurred when the sensemaker revised a likelihood based on his review of the Al Qaeda manual. *Revising* also occurred when the sensemaker revised likelihoods associated with students or non-students causing flight zone violations, based on his analysis of flight paths. *Updating* occurred only in the final cycle of sensemaking, when posteriors from the previous cycle became the priors that were then updated with likelihoods of evidence from FBI interviews.

All three types of re-framing are clearly applicable and important in real-life sensemaking. However, ICArUS challenge problems are constrained to the last type of re-framing, i.e., *Updating*, for three reasons. First, as described in Section 6.1, likelihoods are given to subjects so they are not estimating and re-estimating (i.e., *Revising*) their own likelihoods. Second, as described in Section 6.2, hypotheses are given to subjects so they are not generating (i.e., *Abducting*) their own hypotheses. Finally, the need to avoid RASKRs requires that subjects be

prevented from using their expert knowledge to estimate (and re-estimate) likelihoods or generate hypotheses. These processes are governed by RASKRs in real-world sensemaking, so they simply cannot be tested in challenge problems that avoid RASKRs.

Needless to say, any or all of these constraints (imposed on ***likelihoods***, ***hypotheses***, and ***evidence***) might be relaxed in extending and applying the Octaloop beyond ICArUS challenge problems. Several ideas for such extensions are outlined in Section 7 of this document.

6.4 Three Functions of Sensemaking

In summary of how the Octaloop applies to ICArUS challenge problems, it is useful to highlight three sets of steps in the Bayesian framework. These three sets of steps model three functions of sensemaking, namely: *Inferencing*, *Decision-Making*, and *Foraging*.

First and foremost, sensemaking is a process of forward and backward *Inferencing*. The forward inferences include estimating likelihoods of the form P(e|H) for evidence (e) given hypotheses (H). The backward inferences involve aggregating priors and likelihoods to obtain posteriors of the form P(H|e). Because likelihoods are given to subjects in human experiments, ICArUS challenge problems are focused on the backward inferences in a form of re-framing known as Bayesian *Updating* (see Section 6.3). The steps of the Octaloop that model this function are [1], [2], [3], and [4].

The remaining steps of the Octaloop address choices (decisions) made by a sensemaker, based on judgments (inferences) made in earlier steps noted above. These choices greatly complicate the control of experiments performed with ICArUS challenge problems. The reason is that each choice changes the "game state" (context) for future judgments, so each subject will actually be receiving different stimuli depending on their sequence of choices throughout a "game task" (mission). This makes it difficult or impossible for experimenters to compute a meaningful "average" of human sensemaking performance across subjects, as required by the ICArUS BAA.

For that reason, only selected missions of the ICArUS challenge problems involve choices by participants, and the options for choices are extremely constrained to maintain experimental control. These missions address choices of two types, namely: *Decision-Making* and *Foraging*. *Decision-Making* refers to operational choices of the sort that are not usually made by intelligence analysts. However analysts do analyze possible courses of action, and advise operational decision-makers, so this function of sensemaking is captured in several missions of the ICArUS challenge problems. The relevant steps of the Octaloop are [5] and [6].

Another type of choice more often made by intelligence analysts involves the active collection or passive attention to further information about a situation. In order to address this function, known as *Foraging*, several missions of ICArUS challenge problems relax the constraint of Section 6.2 on giving evidence to participants. For these missions, subjects must choose among several intelligence sources or different areas where intelligence may be obtained, so not all subjects will process the same evidence in the same order. The relevant steps of the Octaloop are [7] and [8].

As outlined above, ICArUS challenge problems address all eight steps of the Octaloop – albeit each to a greater or lesser extent in order to satisfy practical constraints on human experiments and meet BAA requirements. This is done by designing suites of "game tasks", where each task is a simulated mission that challenges one or more of the three sensemaking functions, namely: *Inferencing* [1]-[4], *Decision-Making* [5]-[6], and *Foraging* [7]-[8]. Details of all missions are documented elsewhere for the Phase 1 (Burns, Greenwald, & Fine, 2014) and Phase 2 (Burns, 2014) challenge problems.

6.5 Key Insights on Biases

An important objective of the ICArUS program is to gain new insights into cognitive biases and how they might be overcome. Thus in accordance with the ICArUS BAA, challenge problems are designed to address eight specific heuristics and biases, namely: *Anchoring and Adjustment*, *Availability*, *Change Blindness*, *Confirmation Bias*, *Persistence of Discredited Evidence*, *Probability Matching*, *Representativeness*, and *Satisfaction of Search*. All eight are commonly considered to adversely affect intelligence analysts. Each is a heuristic strategy or the resulting effect (i.e., bias) thereof, relative to normative (Bayesian) standards.

A difficult problem faced in experimental design is to formally define each bias in the context of intelligence sensemaking, and then actually compute normative-Bayesian solutions for the challenge problems – as needed for measuring the existence and magnitude of each individual bias. Some insights gained from this effort are offered below.

6.5.1 An Insight in Hindsight

One insight from the design effort was that some of the so-called biases may actually be normative behaviors, at least when scientists properly consider the context in which judgments and decisions are being made – including the existence of uncertainty and practical constraints on time and effort needed to make decisions and take actions. In short, human heuristics can often be seen as bounded-Bayesian strategies, where any so-called "bias" lies more in the minds of scientists who fail to consider the practical constraints and natural bounds of participants, rather than in the minds of participants who must deal with those bounds and constraints.

For instance, *Probability Matching* has been shown to be an optimal strategy (Burns & Demaree, 2009) when the decision maker is uncertain about the state of the world, i.e., when he or she must earn (exploit) utility from the environment and at the same time learn (explore) the parameters (i.e., probabilities and utilities) of the environment. Similarly, *Change Blindness* and *Satisfaction of Search* are obvious biases only when scientists do not consider the practical benefits of change detection or search completion, as well as the costs in terms of time and effort. For example, if a potential change is non-consequential, then a bounded-Bayesian would be optimal to ignore it. Similarly, if exhaustive search is not expected to be worth the costs, then

a bounded-Bayesian would be optimal to terminate the search when the costs outweigh expected benefits.

Of course this is not to say that biases do not exist. But it does appear that some so-called biases may represent bias in the mind of an observer (of the actor's behavior) rather than bias in the actor's behavior itself. This is especially true for *Confirmation Bias*, which is perhaps the most celebrated bias in circles of intelligence analysis (Heuer, 1999)

An example arises when an actor's "confirming" judgment or decision is observed by others to be "incorrect", from the *deterministic* perspective of what happened *after the fact*. Clearly the causes of bad outcomes are important to assess. But there is always uncertainty before a judgment or decision is made, so the only way to ascertain whether the actor was biased or not is to look at all the relevant probabilities and utilities – as they were known (or could possibly have been known) before the judgment or decision was made. Sometimes low probability events do occur, and sometimes high probability events do not occur. So it is the foresight probability P rather than hindsight probability 1 or 0 that must be considered in assessing the goodness or bias of any judgment or decision that preceded the observed outcome.

6.5.2 The Benefits of Biases

Early research on *Confirmation Bias* (Wason & Johnson-Laird, 1972), done with contrived problems in deterministic settings, found that humans tended to seek information that would confirm their favored hypothesis. In that context the behavior was biased because instead participants should have tried to refute their favored hypothesis. However, subsequent research (Klayman & Ha, 1987) in a probabilistic context more relevant to real-world situations has demonstrated that a "positive test strategy" (i.e., seeking information about the most probable hypothesis) is actually optimal – in the sense of maximizing the expected gain of information.

This same result was obtained in design of ICArUS challenge problems. Detailed calculations were performed for prototypical problems of intelligence collection, using a range of realistic values for the sensor parameters known as "hit", "miss", "false alarm", and "correct rejection" rates per signal detection theory. These Bayesian analyses showed that a positive test strategy was always optimal (i.e., maximizing the information gain), such that the only so-called "bias" would be if a participant in experiments did not choose to obtain further information about the most probable hypothesis – and even that behavior (much like *Probability Matching*, discussed above) may be optimal given second-order uncertainty (i.e., the probabilities of probabilities).

6.5.3 Confirming Conservatism

Another flavor of *Confirmation Bias* (Nickerson, 1998) deals with *weighing* evidence rather than *seeking* evidence. For example, a study of *Confirmation Bias* in the context of intelligence analysis (Lehner, et al., 2008) presented 60 items of evidence to participants in four stages (i.e., 15 items per stage). The study measured subjective confidence in each of three competing

hypotheses {H1, H2, H3}, after each set of 15 evidence items was received. The study also measured subjective "diagnosticity" with respect to each hypothesis, for each individual item of evidence, on a scale from -2 (strongly refutes) to +2 (strongly supports). The study found that judgments of diagnosticity were correlated with confidence in each hypothesis. For example, participants who favored H1 reported that a supporting item of evidence was more diagnostic in supporting H1, compared to participants who favored H2 or H3. Similarly, participants who favored H1 reported that a refuting item of evidence was less diagnostic in refuting H1, compared to participants who favored H2 or H3.

Although interesting, it is not clear to what extent this behavior is actually a *Confirmation Bias* in the normative sense of deviating from Bayesian standards. One issue is that experimenters used judgments of diagnosticity to measure bias, yet functionally these judgments were only inputs to the participants' overall task of aggregating evidence (and associated judgments of diagnosticity) in order to assess confidence across the three hypotheses. The experimenters did not compute a normative solution for confidence across hypotheses after each item of evidence (or after each stage of 15 evidence items), so there was no basis for concluding that participants were biased in that regard. Also, with 60 items of evidence, many items might be rendered non-diagnostic after receiving dependent items, even if each item was diagnostic when considered individually and independently.

In short, the experimental conclusions were based on human judgments of diagnosticity for individual items of evidence, rather than human judgments of confidence in competing hypotheses. Also the measure of bias was computed between different sets of subjective judgments, rather than between subjective judgments and normative standards. To the extent that there was a measured *Confirmation Bias*, it was a bias in estimating individual likelihoods (i.e., diagnosticity of evidence) – and other biases in aggregating likelihoods to compute posteriors may compensate for or even reverse the overall direction of this bias. Numerous experiments on Bayesian inference have shown the overall bias is almost always the opposite of *Confirmation Bias* – as humans are *Conservative* (Edwards, 1982), i.e., *Regressive* toward uniform distributions across hypotheses. For example, if the Bayesian posteriors are {0.99, 0.01}, then humans typically report numbers closer to {0.50, 0.50} such as {0.90, 0.10}. The same result was found in ICArUS experiments – and in fact ***Conservatism*** was by far the most common bias measured across all tasks of Phase 1 and Phase 2 challenge problems.

Of course in some cases the tendency to confirm what one thinks may appear to be a non-normative behavior. One example surfaced in analyzing the story of Section 5, where the sensemaker's desire to confirm his hunch about students versus non-students delayed his search for evidence about a less probable but more consequential hypothesis regarding Al Qaeda. However, the only way to tell if and where a *Confirmation Bias* or any other bias actually exists is by measuring the underlying probabilities (in weighing evidence) and utilities (in seeking evidence). This is rarely done in the real-world of intelligence. But it could be done, by requiring analysts to report their beliefs in the form of numerical probabilities and utilities. That approach is common practice in risk analysis and management of hazardous industries, and it can also be applied to improve the rigor of intelligence analysis (Garrick, et al., 2004; Lehner, et al., 2012; Friedman & Zeckhauser, 2014). A practical idea along these lines is to develop a new structured analytic technique as outlined below.

6.5.4 Substitution of a Structured Technique

A final insight obtained across all ICArUS experiments is that, when subjects are biased, it is usually because they simply do not know the proper way to solve the problem. In that case the bias can be characterized as ***Substitution*** (Kahneman, 2011) of a familiar but improper strategy for the proper Bayesian strategy. Of course subjects themselves do not know they are *Substituting*, because they have not been taught the Bayesian strategy! So the obvious idea for improving intelligence lies not in "de-biasing", but rather just teaching Bayesian reasoning in the first place.

More specifically, the idea is to develop a "Structured Analytic Technique" to support Bayesian analysis in accordance with the Octaloop, as an alternative to the numerous ad hoc techniques that have been proposed to reduce biases. This would avoid the popular (but questionable) pursuit of cataloging biases, and advance the practice of intelligence by teaching the rigors of Bayesian reasoning needed to avoid biases in the first place (regardless of how any biases might be catalogued). More details are provided in Section 7.1 below.

7 Transition

The Octaloop was developed to formalize the study of sensemaking in laboratory challenge problems. However, the same framework holds promise for transition of ICArUS to real-world intelligence in three areas of application, namely: analytic techniques (Section 7.1), analyst training (Section 7.2), and automated tools (Section 7.3).

These applications would all require further effort, because they are beyond the scope of research funded under the ICArUS BAA. Therefore the suggestions below are only ideas about what might be done with further investment. Although cost estimates are not developed here, the three areas are listed roughly in order of increasing investment, from techniques, to training (of techniques), to tools. Likewise, the ideas range from specific to speculative in moving from techniques, to training, to tools.

In that light, the HELP technique discussed in Section 7.1 may be the most promising opportunity for short-term, low-cost, and high-yield transition of ICArUS to the Intelligence Community.

7.1 Technique to HELP Perform Bayesian Reasoning

The Octaloop is a Bayesian-computational framework for rigorous *analysis* of sensemaking. But the same Bayesian framework can also be used for rigorous *synthesis* of intelligence. In other words, the Octaloop can be re-cast in the form of a so-called "Structured Analytic Technique" (SAT), similar to SATs (Beebe & Pherson, 2012) such as the Analysis of Competing Hypotheses (ACH) proposed by Heuer (1999).

In fact ACH is basically a qualitative approach to performing the quantitative analysis modeled more formally by steps [1]-[4] of the Octaloop. As such, ACH simplifies some important aspects of the approach, and this has pros and cons. A pro is that the simplifications make ACH more approachable to analysts who do not have a background in quantitative sciences. A con is that the simplifications can mislead analysts, by not providing the requisite structure to support principled reasoning in accordance with Bayesian standards. Four specific examples of this con are identified in the subsections below.

7.1.1 Hypotheses

Step 1 of ACH says: "*Identify the possible hypotheses to be considered*", which is similar to step [2] of the Octaloop. But as step 1 of ACH, this tends to obscure the fact that hypotheses are always generated in the context of evidence, simply because hypotheses are being generated to explain evidence. In the Octaloop, step [2] Generating Hypotheses is preceded by step [1] Isolating Evidence. This is to highlight the causal connections between hypotheses and evidence in likelihoods of the form P(e|H). Likelihoods are naturally represented in human minds via the

strengths of associative memories (between e and H), and these associative strengths are ultimately how hypotheses are generated in the first place.

For example, a SAT based on the Octaloop might advise analysts to imagine other (unobserved) evidence, in order to create a more comprehensive set of hypotheses than they would naturally generate on the basis of observed evidence. This "what if" technique goes beyond the guidance of existing SATs such as "Brainstorming" or "Starbursting" (i.e., simply asking "who, what, when, where, how, why?"), which are not guided by a science of how humans actually generate hypotheses – i.e., via likelihoods that represent the strengths of associative memories.

A more important limitation of ACH, and of SATs for generating and evaluating hypotheses, is a lack of guidance on what constitutes a well-formed set of competing hypotheses. Specifically, Bayesian reasoning requires that the hypotheses be mutually exclusive, like A and ~A; and exhaustive, like P(A) + P(~A) = 1. Analysts might be more inclined to impose these proper constraints on competing hypotheses if they had a SAT based on the rigors of the Octaloop.

7.1.2 Evidence

Step 2 of ACH says: *"Make a list of significant evidence and arguments for and against each hypothesis"*. Step 2 then goes on to highlight the importance of *assumptions* as well as *evidence* and *arguments*. However, these terms are not clearly distinguished by ACH. For example, items of *evidence* as well as *assumptions* are listed as rows of a matrix, in the next step of ACH (see Section 7.1.3). But only some assumptions are treated in this manner, akin to *evidence*, whereas other assumptions affect *arguments* for the values assigned to cells in the matrix.

In the Octaloop, assumptions are not treated as evidence. Instead only factual information is treated as evidence, with the understanding that these facts may actually be uncertain (and in that respect each item of evidence is also an assumption). The point here is that assumptions based only on background knowledge, which ACH treats the same as evidence, are not categorically equivalent to evidence – so they should not be treated the same as evidence in rows of the matrix. Instead the Octaloop models this background knowledge with a prior probability for each hypothesis.

With respect to *arguments* in ACH, it is not clear if the term refers to some evidence; or a hypothesis; or likelihood of some evidence (given a hypothesis); or confidence in a hypothesis (given some evidence); or some combination thereof (or something else). In the Octaloop, arguments are akin to the causal "reasons" (i.e., bases) for numerical likelihoods and priors. As an example, in the story of Section 5 the sensemaker could think of reasons why students would cause flight zone violations, and these reasons affected his likelihood estimate. Similarly, he could think of reasons why Al Qaeda might be active shortly after 9/11, and those reasons affected his prior.

7.1.3 Likelihoods

Step 3 of ACH says: *"Prepare a matrix with hypotheses across the top and evidence down the side. Analyze the 'diagnosticity' of evidence and arguments – that is, identify which items are most helpful in judging the relative likelihood of the hypotheses."*

This is a key step in which there are several problems. One problem mentioned above is that the matrix does not model priors of the form P(H), and instead treats assumptions as evidence. Another problem is that the method confuses likelihoods of the form P(e|H) with posteriors of the form P(H|e), by saying to analyze *"diagnosticity"* and judge the *"relative likelihood of the hypotheses"*. This suggests that the analyst should estimate P(H|e) directly for each cell in a row, i.e., by estimating the probability of each hypothesis taking the evidence to be true. But actually the input estimates needed for Bayesian inference are likelihoods of the form P(e|H), which are later aggregated to estimate posteriors of the form P(H|e) as outputs (see Section 7.1.4). So instead step 3 should advise analysts to estimate likelihoods of the form P(e|H).

Similarly, the method says to assess "diagnosticity" by taking an item of evidence and asking whether it is *"consistent with, inconsistent with, or irrelevant to each hypothesis"*. But this notion of "consistency" is not the same as "causality", and likelihoods (needed as input to the matrix) are probabilities of evidential *effects* (e) assuming hypothetical *causes* (H), i.e., P(e|H). Also, besides P(e|H), a Bayesian would address P(~e|H) to ensure that P(e|H) + P(~e|H) = 1.

Another limitation of the matrix is that the values for diagnosticity (or consistency) are indicated by symbols such as "+" and "-". The method says that these symbols can be replaced by numbers (e.g., probabilities), if they are known, but does not say how the symbols would be mapped to those numbers (i.e., what does "+" or "-" mean, numerically?), or how numbers might be developed (if they are not known) in order to improve the rigor of analysis.

7.1.4 Posteriors

Step 4 of ACH is to review and refine the matrix. Then step 5 says: *"Draw tentative conclusions about the relative likelihood of each hypothesis. Proceed by trying to disprove hypotheses rather than prove them."* In this step the term "likelihood" is being used to mean "probability", and refers to posteriors of the form P(H|e) rather than Bayesian likelihoods of the form P(e|H). Nevertheless, the purpose of the step is to perform a qualitative aggregation of likelihoods down columns of the matrix in order to estimate posteriors not represented in the matrix.

The main problem is that ACH is not clear about the logic for aggregating values, working down each column of the matrix. For example, it says: *"The pluses... are far less significant [than the minuses]*." This can be confusing because the symbols "+" and "-" are usually considered additive opposites and do not imply any magnitude or significance. Also, as we know from analyzing the story in Section 5, a sensemaker often wants to know what most probably did (+) cause the evidence, not what most probably did not (-) cause the evidence. So in that sense the pluses may be quite significant. Simply telling analysts to focus on the "-" cells of the matrix,

without a more formal and transparent method for aggregating symbols like pluses and minuses down the columns, seems to fall short of the structure that is needed for an effective Structured Analytic Technique.

Another problem with the matrix of ACH is that it fails to account for dependencies between different items of evidence (and between different assumptions that are treated as evidence in rows of the matrix). For example, in the story of Section 5, FBI interviews were presumably different and diverse from the counterintelligence reports considered earlier. So the two sources could reasonably be assumed independent. But often there will be dependencies between sources of evidence, and sometimes there will be complete dependency such that one item of evidence (or prior assumption) renders a second item of evidence (or prior assumption) completely non-diagnostic with respect to the competing hypotheses. These effects are not addressed by the matrix and method of ACH, where values (akin to likelihoods or priors) are entered across a row without regard to other rows (above or below) on which that row may be dependent.

This problem is of particular concern when *assumptions* are treated as *evidence* in the rows of the matrix. As noted above, assumptions actually reflect priors (i.e., in the absence of evidence) for hypotheses, not likelihoods of evidence given hypotheses. These priors are always estimated in the context of a world view held by the analyst, and that world view affects all assumptions – so it represents a major dependency between rows of assumptions.

For example, using ACH it is tempting for an analyst to list as many *assumptions* (in rows of the matrix) as possible, especially when there is little *evidence* (in rows of the matrix) to use in the analysis of competing hypotheses. Using the Octaloop, there is only one prior probability for each hypothesis, and it reflects all the assumptions of an analyst's world view. This helps avoid the pitfalls of treating assumptions as evidence.

7.1.5 HELP

Despite the limitations noted above, ACH was a significant step in advancing the rigor of intelligence analysis. Intelligence analysts often refuse or resist quantitative approaches, and the principles of Bayesian reasoning as applied to intelligence analysis are not trivial to teach (Grabo, 2004; Schweitzer, 1976; Fisk, 1972; Zlotnick, 1970; Edwards, et al., 1968; Edwards & Phillips, 1964). But these difficulties are precisely why the Intelligence Community should invest in formal techniques with more rigor than current SATs – of which ACH is actually one of the most rigorous, relative to other SATs that have no apparent science as a basis for their structure.

Previous research on Bayesian reasoning by unaided humans has uncovered four classes of errors (Burns, 2006) that appear to be ubiquitous, namely: (1) failure to generate a mutually exclusive and exhaustive set of hypotheses, (2) failure to distinguish assumptions from evidence, (3) failure to distinguish likelihoods from posteriors, and the consequent failure to correctly estimate causal likelihoods, (4) failure to properly aggregate likelihoods and priors in computing posteriors, including failure to consider conditional dependencies between items of evidence.

Notice that ACH itself exhibits all four of these limitations, as identified in Sections 7.1.1, 7.1.2, 7.1.3, and 7.1.4 above. This suggests that ACH does not provide the structure needed to help humans reason rigorously as Bayesians.

An alternative would be to develop a new SAT with the proper structure. This SAT might be dubbed **HELP**, to highlight key concepts of Bayesian inference and address associated errors outlined above, as follows:

Hypotheses
Evidence
Likelihoods
Priors and **P**osteriors.

Bayesian HELP would be based on the Octaloop, especially steps [1]-[4] of *Inferencing* and steps [7]-[8] of *Foraging*. The treatment of *Inferencing* would be more formal than ACH, addressing the issues noted in Sections 7.1.1, 7.1.2, 7.1.3, and 7.1.4. The treatment of *Foraging* would go beyond ACH and other SATs in addressing what evidence an analyst should attempt to acquire, based on previous inferencing, for input to further inferencing.

After the investment needed to develop and document HELP in the form of a Structured Analytic Technique, further investment would be needed for training analysts in use of the technique. Here, like the Octaloop itself, the present document offers a practical approach – namely stories, discussed below.

7.2 Training of Critical Thinking with Bayesian HELP

Storytelling is a powerful device, widely used for organizational learning and personnel training. But a problem with stories (which contributes to their popularity) is that narratives leave much to the imagination. Often this is not acknowledged by writers and readers, as in the story that Klein, et al. (2007) told to support their data-frame theory of sensemaking. A more formal analysis of this same story, using the Octaloop, uncovered numerous omissions for which we as readers needed to make assumptions in order to compute solutions.

In effect, Section 5 was an exercise in "critical thinking", where each cycle in the story was analyzed to identify all categorical concepts (hypotheses, evidence, likelihoods, priors, and posteriors) and numerical values (probabilities) needed to formalize sensemaking. As such, materials similar to Section 5 might be used in classroom exercises to teach Critical Thinking Skills and Bayesian reasoning. Additionally, more analytical stories could be dissected and documented in the same manner, for use in case-based teaching/learning.

This approach to training would require that the Octaloop first be used to develop HELP as a Structured Analytic Technique (Section 7.1). As-is, materials in the present document were developed for ICArUS purposes and not intended for classroom training or any other application

beyond the scope of ICArUS itself. A suitable document for training purposes would first introduce HELP in the form of a step-by-step procedure for analysis, and then demonstrate HELP using analytical stories like the example in Section 5.

Notice that this approach would not employ ICArUS challenge problems themselves, and would not require any software like that used for performing ICArUS experiments. Instead the approach would rely on the Bayesian-computational basis used to design ICArUS challenge problems, re-cast in the form of a Structured Analytic Technique (HELP) – along with stories as case studies for engaging analysts in applying the technique to examples of real-world intelligence. Training exercises of this sort would stimulate discussions of important issues in real-world analysis that go far beyond the laboratory constraints of ICArUS challenge problems, and these issues would not be addressed if training were to focus on those challenge problems.

Notice also that this approach to training would not focus on cognitive biases. That is, the training would not be to tell whether an inference or decision should be classified as *Conservative* versus *Confirmation* bias, or whether it was caused by a heuristic of *Anchoring* versus *Availability*, etc. Instead the training would focus on how to perform Bayesian reasoning, with HELP, including all information and assumptions that a Bayesian would need to compute a solution.

Elsewhere the exercise of identifying biases in stories has been adopted, in recent efforts to teach critical thinking using game-based methods at the Mercyhurst University Institute of Intelligence Studies (Richey, 2014). But identifying biases in others is not the same as identifying biases in oneself, and identifying biases is not the same as avoiding or overcoming biases. Plus it is not clear how analysts can identify biases in stories without first being trained in the Bayesian methods by which many of those biases are defined. Instead it seems more useful to use stories for teaching analysts the principles of Bayesian reasoning in the first place, e.g., as structured by the HELP technique.

7.3 Tools for Cognitive Support to Analysts

Sections 7.1 and 7.2 focused on a technique and training (of the technique) to support human analysts in sensemaking. A third area where the Octaloop might be useful is in the design of automated tools for intelligence. The idea here is to offload some step or steps of the Octaloop onto a system that can offer ***cognitive*** support to humans engaged in the cycle of sensemaking.

As described in Section 6, ICArUS challenge problems actually do this to some extent in order to facilitate human experiments. Likewise, real-world systems and sensors are often designed to perform functions such as [8] Discriminating Evidence, e.g., in a system that alerts a user to some observed activity (i.e., ***evidence***); or [3] Estimating Likelihoods, e.g., in a sensor that returns data with some reliability (i.e., ***likelihood***). But the uses of such systems/sensors by human beings pose engineering problems far beyond the "point solutions" computed by the systems/sensors, because these tools are only useful if they are understood and accepted by

human users who are engaged in the complete cycle of sensemaking. That is, useful tools must provide ***cognitive*** support to analysts.

Software developed for ICArUS experiments simulates such sensors/systems and provides input to human subjects at steps [8], [1], [2], and [3] of the Octaloop. Steps [4], [5], [6], and [7] are then performed by human subjects. ICArUS experiments have measured and modeled various cognitive biases at steps [4], [5], [6], and [7], and those biases represent opportunities for designing analytical support systems. One example is the colored calculator dubbed "Bayesian Boxes" (Burns, 2006), which offers an intuitive visualization of Bayes Rule, and has been proven (Burns, 2007) to reduce biases such as *Anchoring* and *Averaging* at step [4] of the Octaloop. This system has also been used for real-world training of Bayesian inference, and could be used for "drawing pictures" while "telling stories" in training Bayesian HELP (see Section 7.2).

Other systems might be designed to help humans at steps [5], [6], [7]. As discovered in ICArUS challenge problem design, Bayesian solutions are especially difficult to compute at step [7]. The calculations require predictions (based on likelihoods) about what evidence will be received, and projections of the updated (posterior) probabilities across hypotheses – for all possible outcomes (of collecting information) from all feasible actions (for collecting information). These computational challenges at step [7] represent practical opportunities for providing automated support to analysts.

Additional applications to system design might address steps [8], [1], [2], and [3]. At these steps of the Octaloop, ICArUS software implements prototype systems that provide point solutions to human users. The practical value of these prototypes is that they demonstrate what functions might usefully be performed by systems to support humans in the complete cycle of sensemaking. In particular, because sensemaking hinges on ***likelihoods***, the challenge problem software simulates various intelligence systems/sensors and associated ***likelihoods*** – e.g., for HUMINT, IMINT, SIGINT, OSINT, etc. – as needed for human subjects (and neural models) to analyze competing hypotheses given multiple INTS as evidence.

As a concrete example, Estimating Likelihoods at step [3] in the real world of intelligence is often referred to as "suitability analysis". Geospatial analysts are routinely engaged in analyzing the suitability of terrain or some other feature f (e.g., from IMINT) that constrains the probability of a hypothesized activity. We saw this in the story of Section 5, where *f = airport vectors found in geospatial overlays*, and this feature suggested that flight paths over nuclear power plants would be more "suitable" for students than for experienced pilots.

The corresponding ***likelihoods*** were estimated as $P(s|S,f) = 0.95$ and $P(s|\sim S,f) = 0.10$, where the geospatial feature f served as context to constrain the likelihood that overflight activity would be observed (s) if the culprits were students (S) or non-students (~S). These likelihoods from step [3] were then used in step [4], where the overflight activity was explained by computing confidences across hypotheses – e.g., in a form of "activity-based intelligence" analysis.

Then, in a subsequent cycle of sensemaking, additional ***likelihoods*** were estimated using the FBI interviews (e.g., from HUMINT) and aggregated with earlier ***likelihoods*** to further explain the situation. The point here is that data (e.g., evidence from INT sources) is useful for sensemaking

only if ***likelihoods*** (of evidence, given hypotheses) are either assumed by the human (as in the story of Section 5) or given to the human by another human or "system" (as in ICArUS experiments). Therefore an application for ***cognitive*** support to analysts lies in designing systems to estimate likelihoods that are difficult for humans to estimate themselves.

For instance, task [3] Estimating Likelihoods might be performed by a system that searches large databases and computes historical frequencies of past events (e) in the context of their known causes (H), which can then be taken as measures of likelihoods, P(e|H). Besides such forensic analyses, other systems designed for modeling and simulation of real-world situations might be used to estimate likelihoods in prognostic analyses. This would involve running the simulations in a parametric sampling mode, i.e., obtaining results for different sets of assumed parameters representing different contexts c (like f in the suitability analysis above), in order to compute statistical likelihoods of the form P(e|H,c). The key here lies in identifying the categories of hypotheses H, evidence e, and context c that are relevant to a sensemaker, so that probabilities of the form P(e|H,c) can be computed to provide ***cognitive*** support to analysts.

Opportunities for analytical support systems may also exist at steps [8], [1], and [2] of the Octaloop. But Generating Hypotheses at step [2] hinges on the rich knowledge (RASKR) of human experts. And Isolating Evidence at step [1] appears difficult to automate in a fashion that would support a human expert engaged in adaptive analysis of competing hypotheses. However, step [8] Discriminating Evidence may be amenable to automation by a system that screens large volumes of information beyond what a human could possibly consider. Here again, the key is to provide ***cognitive*** support – so the system would need to anticipate and adapt to features of evidence and classes of hypotheses that are relevant to the sensemaker.

Although speculative, the above ideas illustrate how the Octaloop might guide the design of advanced tools to support sensemaking. The same logic also extends beyond individual humans and systems to organizational structures for team sensemaking, where different humans perform different steps of the sensemaking cycle and share results with each other. In that case the Octaloop might shed light on what the different teammates need from one another, and how they can best function together across steps and cycles of the Octaloop, including interfaces between intelligence analysts and operations personnel.

7.4 Leveraging the Bayesian Research Community

As a final idea for ICArUS transition, the Octaloop might serve as a framework for eliciting and applying contributions from the Bayesian research community. In particular, it has been suggested (NRC, 2010) that decades of Bayesian research are untapped and might be leveraged to improve intelligence analysis, and actually this was the topic of several workshops sponsored in Phase 2 of ICArUS.

But it is not clear how much of this Bayesian research is useful for intelligence analysis, and there remains a large gap between analysts and researchers at gatherings of the two groups. Typically analysts set the stage by presenting examples of their real-world challenges, often in

the form of anecdotal stories like that of Klein, et al. (2007). Then researchers follow by presenting their findings from the laboratory, in the form of theoretical models and experimental results – but on problems that bear little or no resemblance to real-world intelligence. So researchers have not shown how their methods and models can be applied to aspects of the stories told by analysts. And analysts have not shown how aspects of their stories might be addressed by Bayesian tools, training, or techniques.

One reason for this gap is that the two sides are lacking a common framework for understanding the cognitive-computational challenges of intelligence analysis. The Octaloop can help close that gap because it is expressed in formal language (Section 3) consistent with Bayesian methods (Section 4), and because it applies to analytical problems (Section 5) cast in the form of anecdotal stories (Section 2). To illustrate these advantages, the Octaloop has been used to generate specific ideas about how a Bayesian approach can advance intelligence techniques, training, and tools – in Sections 7.1, 7.2, and 7.3, respectively. Those ideas might serve to stimulate further discussions between researchers and analysts.

For instance, the present document or portions thereof could be provided to Bayesian researchers with the challenge of answering the following question:

> *Using the Bayesian framework (Section 4), can you show exactly where and how your Bayesian model or method would enable machine automation (i.e., a tool) or improve human cognition (i.e., a technique or training) in some aspect(s) of the story (Section 5)?*
>
> *Be specific, and provide one or more examples where you actually compute something that applies directly to the story. If necessary, assume any details that are not contained in the narrative but are necessary for you to show how your model or method would apply.*

Analysts could then be asked if they think a researcher's example is relevant, and whether they can envision it being applied beyond the story world to real-world intelligence. Analysts could also be asked to provide additional stories that might be analyzed in the same Bayesian framework and utilized for the same purpose of closing the gap between research and practice.

8 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.

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

Burns, K. (2014). *ICArUS Phase 2 Challenge Problem Design and Test Specification*. MITRE Technical Report, MTR140412.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Friedman, J., & Zeckhauser, R. (2014). Handling and mishandling estimative probability: Likelihood, confidence, and the search for Bin Laden. *Intelligence and National Security.*

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

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

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

Kahneman, D. (2011). *Thinking Fast and Slow*. New York: Farrar, Straus, & Giroux.

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

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

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

Lehner, P., Michelson, A., Adelman, L., & Goodman, A. (2012). Using inferred probabilities to measure the accuracy of imprecise forecasts. *Judgment and Decision Making, 7(6),* 728-740.

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

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

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

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

NRC (2010). National Research Council, Field *Evaluation in the Intelligence and Counterintelligence Context: Workshop Summary*. Washington, DC: National Research Council.

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

Pirolli, P. (2007). *Information Foraging Theory: Adaptive Interaction with Information*. Oxford, UK: Oxford University Press.

Richey, M. (2014). The mind's lie: Games-based learning for critical thinking. *The International Journal of Humanities Education, 12(1),* 1-12.

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

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

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

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

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