Cognitive Assistance in Government
Papers from the AAAI 2015 Fall Symposium

Bayesian HELP: Assisting Inferences in All-Source Intelligence

Kevin Burns
The MITRE Corporation
kburns@mitre.org

Abstract

Abductive inference is a cognitive competence of intelligence analysts, in which alternative hypotheses are generated and evaluated to explain and predict evidence. This paper presents a formal framework for artificial intelligence to assist analysts in making such inferences, based on a Bayesian approach to analyzing hypotheses, evidence, likelihoods, priors, and posteriors (HELP). The approach is applied to a case study of natural intelligence, in order to identify those components of cognitive performance that offer opportunities for assistance. Existing tools and techniques for such assistance are also reviewed, to establish how well they can help analysts overcome well-known heuristics and biases in probabilistic reasoning. The review suggests that currently available techniques do not help, and may even hurt by encouraging some of the most common errors. This finding motivates the recommendation of a structured analytic technique and tools for supporting Bayesian analyses of competing hypotheses, using the formal framework of HELP.

Abductive Inference

All-source intelligence involves generating and evaluating alternative hypotheses to explain and predict evidence obtained from multiple sources. This competence, sometimes called *sensemaking* (IARPA 2010), is how analysts come to understand situations and recommend courses of action.

Previous research has proposed conceptual theories to describe cognitive sensemaking (Klein, Moon, and Hoffman 2006a; 2006b; Klein et al. 2007). But those theories do not specify the cognitive mechanisms in a computational manner, as needed to model and measure the inferences of analysts – and as needed to design artificial intelligence (AI) that can offer cognitive assistance. The present paper instead offers a formal framework using Bayesian principles (Bayes 1763; Fischhoff and Beyth-Marom 1983; Mueller 2009) to mathematically model the cognitive components of sensemaking.

This framework (Burns 2014a; 2005) includes the following components collectively dubbed HELP: *hypotheses*, *evidence*, *likelihoods*, *priors*, and *posteriors*. The hypotheses are possible explanations of actual evidence that has been received or potential evidence that might be received. The likelihoods, priors, and posteriors are each represented by a probability ranging from zero to one. A likelihood, denoted P(e|H), is the probability of some evidence (e) assuming the truth of a hypothesis (H). A prior, denoted P(H), is the probability of a hypothesis in the absence of some evidence. A posterior, denoted P(H|e), is the probability of a hypothesis (H) given some evidence (e).

The basic idea of Bayesian inference is one of updating prior probabilities to compute posterior probabilities, and this applies iteratively. That is, the posterior probability of a hypothesis (after some evidence) becomes the prior probability for that hypothesis in a future update with further evidence. The updating is accomplished using Bayes' Rule, which states that a posterior probability is computed as the normalized product of a prior and likelihood, $P(H_i|e) = P(H_i) * P(e|H_i) / P(e)$. The normalizing factor P(e) is a marginal probability computed as the sum of products $P(H_i) * P(e|H_i)$ over all hypotheses in a set $\{H_i\}$ of mutually exclusive and exhaustive hypotheses.

These components of Bayesian HELP serve to formalize the conceptual notion of a *frame*, which previous authors of a data-frame theory (Klein et al. 2007) have described only vaguely as "a story... map... script... plan... [or other] structure for accounting for the data and guiding the search for more data." Unlike the data-frame theory, Bayesian HELP specifies the computational components of this structure. Also unlike the data-frame theory, which separates data from its frame, a Bayesian frame always includes data (i.e., evidence) as well as other knowledge and beliefs (i.e., hypotheses, likelihoods, priors, and posteriors) by which one makes sense of the data. The reason is that likelihoods are needed for computing confidence in hypotheses, and likelihoods always refer to data (evidence) because a likelihood is the probability of some evidence given a hypothesis.

The components of Bayesian HELP also serve to formalize the conceptual notions of *framing* and *reframing* (Klein et al. 2007), i.e., as processes for computing confidence across a set of hypotheses. In fact there are at least three different types of reframing that can be distinguished as follows: *updating*, *revising*, and *abducting*. In *updating* (described above), new evidence and associated likelihoods are used to update priors and compute posteriors via Bayes' Rule over a fixed set of hypotheses. In *revising*, old likelihoods are replaced by new likelihoods and a previous update is repeated, again over a fixed set of hypotheses. In *abducting*, new hypotheses are generated along with associated priors and likelihoods of evidence, and posteriors are computed over the new set of hypotheses.

Ominous Airplanes

Here the approach, outlined above, is applied to a case study of natural intelligence by a real-world analyst. The story was originally published by Klein et al. (2007) to motivate their conceptual theory of sensemaking. The same story is dissected below to demonstrate the mathematical model of Bayesian HELP – and to illustrate opportunities for AI to computationally assist the cognitive inferences of intelligence analysts. The story involves five cycles of sensemaking, as an all-source analyst acquires evidence and updates his confidence in competing hypotheses. Each numbered cycle below begins with a quoted excerpt from the story, followed by an analysis using HELP.

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 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 there were suddenly a number of reports of small, low-flying planes around these facilities. At face value it appeared 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)."

This first cycle of sensemaking begins as the sensemaker (hereafter denoted M) attends to an item of evidence from counterintelligence, denoted here as s = sudden increase (after 9/11) in reported flight zone violations. M thought this constituted a terrorist threat, so he was generating hypotheses $\{H_i\}$ about possible causes of the evidence s and estimating likelihoods of the form $P(s|H_i)$. In fact mental likelihoods of the form $P(s|H_i)$ would govern which hypotheses are recalled or constructed from long-term memory and represented in working memory as possible explanations of the observed evidence s. The story mentions a hypothesis denoted here as A = Al Qaeda, and suggests there was a strong association between A and s in the mind of M such that P(s|A) was large. Although the story does not say, M would also have generated the hypothesis ~A = Not Al Qaeda, to represent other possible explanations, because he was clearly not certain that the evidence s was caused by A. Finally, besides a set of at least two hypotheses {A, ~A}, and associated likelihoods P(s|A) and P(s|~A), M would also be representing prior probabilities P(A) and P(~A) in his working memory. These priors reflect preconceived beliefs that M brings to the first cycle of sensemaking without regard for the evidence s.

The story does not provide numerical values for any probabilities, and if asked the sensemaker M might even deny that he represented such quantities in his mind. But clearly M is not equally confident in A and ~A, so some measure of relative confidence in these two hypotheses is mentally represented at least implicitly. Similarly, likelihoods of the form P(s|A) and P(s|~A) are represented, at least implicitly, because these likelihoods govern which hypotheses are generated in the first place. For example, the story suggests that P(s|A) is much higher than P(s|~A), because M can think of a reason (i.e., surveillance by terrorists) why A would cause s but does not think of a reason why ~A would cause s.

The point here is twofold: First, hypotheses, evidence, likelihoods, priors, and posteriors (HELP) *may* all be represented in the mind of a sensemaker, at least implicitly and qualitatively, in order for the sensemaker to make sense of what has been sensed (as evidence). Second, the same components of HELP *must* be represented explicitly and quantitatively, in order to rigorously model and measure sensemaking. Therefore, for purposes of quantification here, we can assign numbers that are at least roughly consistent with the story. For example, we might assume P(A) = P(~A) = 0.50 if M's prior confidence was indifferent between A and ~A. However, the events of the story took place soon after the 9/11 attacks when Al Qaeda was prominent in the thoughts of most Americans, so here as rough estimates we might 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 likelihoods 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 corresponding 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) = 0.50 means that s would be a random (i.e., for no causal reason) effect if ~A was true, such that P(s|~A) = P(~s|~A) = 0.50.

Using the priors and likelihoods outlined above, we can complete our Bayesian analysis of how the sensemaker formed his initial belief that s was most probably caused by "bad guys" (A). The posterior is computed as a normalized product of prior and likelihood, for each hypothesis (A and ~A), via Bayes' Rule as follows: P(A|s) = P(A) * P(s|A) / P(s); P(~A|s) = P(~A) * P(s|~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(~A) * P(s|~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 that Al Qaeda's surveillance activities are the most probable explanation of the evidence from counterintelligence.

2. Reviewing Their Tactics

> "Major A. S. obtained access to the Al Qaeda tactics manual, which instructed Al Qaeda members not to bring attention to themselves. This piece of information helped him to begin to form 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 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."

Based on his posterior beliefs after assessing the evidence s, M would have formed expectations about further information that might be obtained and assessed next. Those expectations would affect whether he would seek more information (or not), and where he would seek to obtain it. The story tells us that M obtained access to the Al Qaeda manual, so apparently he expected it would say something that would shed light on the likelihood P(s|A). Although the story does not say, it is reasonable to assume that M expected the manual would provide some information that confirms his suspicions about A, simply because at this point A was the most probable hypothesis. In that light M must have been surprised by what he read, because it was a violation of his expectations. More specifically, M learned that Al Qaeda members are instructed not to bring attention to themselves, and this affected his estimate of the likelihood P(s|A).

For example, we might assume that after reading the Al Qaeda 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 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 of 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). So here we find a form of reframing that involves *revising* likelihoods and associated posteriors across a set of hypotheses {A, ~A}. This revising is the first of three fundamentally different types of reframing that are found in the story, and the other two types will be highlighted later when they occur.

As a result of revising likelihoods and posteriors, 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 began to wonder who, if not Al Qaeda, is likely to break the rules and cause the observed evidence s. Eventually M generated a new 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 think deeper about the hypothesis ~A, in light of the evidence s. In doing so it appears that M was motivated by two things. First, he now thought ~A was the most probable hypothesis. Second, his likelihoods for this most probable hypothesis were P(s|~A) = 0.50 and P(~s|~A) = 0.50, so M 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 *not* responsible for the overflight activity, but he still had no clue as to who *was* responsible – and apparently he felt a strong need to establish who was responsible.

3. Abducting a Reason

> "He thought 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 fly low patterns, from east to west, from airport to airport."

Motivated by his desire to find a causal reason for the evidence s, M initiated this third cycle of sensemaking without the introduction of any new information. That is, M was generating hypotheses about who might be responsible for s, after realizing that Al Qaeda (A) is probably not responsible.

The result is a new hypothesis S = Student pilots (and not Al Qaeda), based on a strong association between S and s in M's mind, which reflects a reason for why S would cause s. That is, based on M's expertise as a pilot, he thinks P(s|S) is high because he knows why students would be likely to fly over nuclear power plants. Numerically, we might assume P(s|S) = 0.90 because students have a reason for causing s, and 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). Also at this point M's reframing involves *abducting* hypotheses and associated likelihoods of those hypotheses (Burns 2015). This is much like the initial framing we saw in the first cycle of sensemaking, and it is clearly more complex than the *revising* (over a fixed set of hypotheses) that we saw in the second cycle.

To complete the analysis of this third cycle, 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 the present cycle we have P(s|S) = 0.90 and P(s|~S) = 0.50. Using Bayes' Rule to compute the posteriors yields: 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.

4. Collecting More Data

"It took Major A. S. about 3 weeks to do his assessment. He found all relevant message traffic by searching databases for about 3 days. He picked 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 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."

As in the second cycle of sensemaking, where M thought to consult the Al Qaeda manual, his beliefs here at the start of the fourth cycle led him to seek further information that might better distinguish the cause (A, S, or ~S) of evidence s. The story does not say why M chose to examine flight paths. But like his earlier decision to read the Al Qaeda manual, it is reasonable to assume that he expected a flight path analysis would confirm his suspicions about the most likely hypothesis (S).

M's assessment of flight paths was a form of suitability analysis, which is typically performed by geospatial analysts to establish whether features of terrain are likely to be suitable for some hypothesized activity. In this case M found that vectors through restricted zones had airports on either end, and the story says this added support to his hypothesis (S). But actually M's findings first affected his estimates of likelihoods, which in turn affected his posterior 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 reframing here is a *revising* of likelihoods and associated posteriors, similar to the revising that we saw in the second cycle where M decreased his estimate for P(s|A) after reading the Al Qaeda manual.

For example, based on his geospatial analysis, we might assume M increased P(s|S) from 0.90 to 0.95 and decreased 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 make these paths quite suitable for lost students. The decrease in P(s|~S) comes from the finding of other flight paths that would be more suitable for experienced pilots.

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 of 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, and M is even more certain than before that the most probable explanation for the overflight activity is student pilots (who are not members of Al Qaeda).

5. Concluding "It's Students"

"He also checked 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 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 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."

An interesting aspect of this story is that M chose to spend days or weeks on the flight path analysis, which would only help distinguish S from ~S, before checking the FBI records. The FBI records would help distinguish A from ~A, and a threat of Al Qaeda activity was M's primary concern at the start of the story. But later in the story it appears that M's priority for further analysis was to establish who *did* cause s (which he suspected was S) rather than who did *not* cause s.

Some might characterize this behavior as a confirmation bias (Nickerson 1998), because M first chose 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 in fact M's behavior may actually be optimal from an information foraging perspective (Pirolli 2007), because a "positive test strategy" (Klayman and Ha 1987) has been shown to maximize the expected gain in information for prototypical situations of intelligence collection (Burns 2014a; 2014b).

Also, if M's objective was to recommend some policy action to mitigate flight zone violations, then he would want and need to know who are the culprits rather than who are not the culprits. Thus like the earlier instances where M chose to obtain evidence that he expected would support his favored hypothesis, it is not clear whether M's confirmation preference is actually a confirmation bias (relative to Bayesian standards). An answer to that question would require that more parameters of the situation be identified and quantified, beyond what can be done here based on the narrative of Klein et al. (2007).

In any case, the new evidence obtained in this fifth and final cycle of sensemaking is: n = no nefarious pilots identified in the FBI interviews. The associated likelihoods are probabilities of n, conditional on each hypothesis {A, S, ~S}, but also conditional on the previous evidence s. Because n comes from a different and diverse source of intelligence than the evidence s from counterintelligence, we can assume n and s are independent such that the likelihoods of n are conditional 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, we 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) = P(~n|S) = 0.50, and also for non-students we have P(n|~S) = P(~n|~S) = 0.50.

Thus the three likelihoods for n are: P(n|A) = 0.01, P(n|S) = 0.50, and P(n|~S) = 0.50, and Bayes' Rule is used to update the posteriors computed in the previous cycle of sensemaking. Those posteriors become priors in the present cycle as follows: 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 not explained by Al Qaeda activity, P(A|n,s) = 0.001. He is also pretty sure that the evidence is explained by activities of student pilots following visual flight rules, P(S|n,s) = 0.90.

Notice the nature of reframing here in this final cycle is one of *updating* confidence in each hypothesis, over a fixed set of hypotheses, based on likelihoods of the new evidence. This updating is different from the *abducting* that we saw in the first and third cycles, respectively, because here no new hypotheses are generated. This updating is also different from the *revising* that we saw in the second and fourth cycles, because here the new likelihoods are used to augment previous likelihoods in an iterative Bayesian update, rather than to replace old likelihoods and repeat an old update. In iterative updating, posteriors from the previous update become priors for the present update.

HELP for Analysts

The above analysis illustrates how HELP can mathematically model an analyst's *sensemaking*, even if the analyst was not consciously attempting to compute anything in his or her mind. The analysis also illustrates how HELP can usefully mitigate cognitive heuristics and biases that may arise in unaided sensemaking, i.e., by providing intelligence analysts with a principled framework for Bayesian analyses of competing hypotheses.

Thus the formal framework is useful in two ways. First, HELP can be used by analysts as a structured analytic *technique* (SAT) for performing Bayesian analyses of competing hypotheses. Second, HELP can be used by engineers for designing AI *tools* to assist analysts in performing Bayesian analyses of competing hypotheses. Importantly, these two uses go hand in hand because the utility of any tool to assist analysts lies in its capability to work in concert with analysts. The two uses are discussed in subsections below.

Techniques for Improving Inferences

Currently there exist dozens of SATs for use by intelligence analysts (Beebe and Pherson 2012). But only one SAT addresses the analytic problem of abductive inference, which involves generating and evaluating alternative hypotheses to explain and predict evidence. That one SAT is a technique called analysis of competing hypotheses (ACH), which is intended to mitigate cognitive biases, especially confirmation bias (Heuer 1999).

Unfortunately ACH does not adequately address cognitive biases, and in fact it appears to encourage a number of common biases. This conclusion is based on a detailed review (Burns 2014a), which assessed the extent to which ACH addressed four classes of errors most commonly found in unaided inferences (Burns 2006), namely:

1. Failure to generate a mutually exclusive and exhaustive set of hypotheses.
2. Failure to distinguish assumptions and arguments 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.

The review found that ACH actually encouraged these errors, because ACH avoids numerical probabilities and instead advises analysts to evaluate a qualitative notion of *diagnosticity* – using a matrix of hypotheses (columns) versus assumptions and evidence (rows). The diagnosticity defined by ACH does not adequately distinguish between likelihoods and posteriors, and the matrix of ACH confounds assumptions, arguments, and evidence in its rows.

HELP differs by using *probabilities* and by properly distinguishing between likelihoods and posteriors. HELP also distinguishes evidence from prior assumptions, and distinguishes evidence from arguments that support likelihood estimates. HELP does so in the following four steps, which were illustrated in each cycle of the sensemaking story about *Ominous Airplanes*:

1. Hypothesis generation
2. Evidence isolation
3. Likelihood estimation
4. Posterior aggregation.

These four steps of HELP address the four classes of errors denoted above by the corresponding numbers 1, 2, 3, and 4.

Tools for Assisting Analysts

As described above, the steps and supporting principles of HELP provide a structured analytic *technique* for performing Bayesian analyses of competing hypotheses. The same steps of HELP also offer a computational basis for designing AI *tools* that can provide cognitive assistance to analysts.

A basic tool might serve simply as a bookkeeping system, to help analysts organize the hypotheses that they generate; and the evidence that they isolate; and the likelihoods that they estimate; and the posteriors that they aggregate. The benefit would come from the system's structuring of these Bayesian entities in a way that helps humans avoid the four common classes of errors outlined above. A better tool might automate any one or more of the four steps, in order to further assist the analyst's inferences.

For instance, the HELP step of likelihood estimation might be supported by a forensic system that searches large databases and computes conditional probabilities of the form P(e|H,c), based on historical frequencies of past events (e) in the context (c) of their known causes (H). Alternatively, a prognostic system might run large numbers of simulations in a parametric sampling mode, obtaining results for different sets of assumed parameters representing different contexts c in order to compute probabilities of the form P(e|H,c). However, the key to such analyses (forensic or prognostic) would be for the AI to discover or be given (via user input) the categorical hypotheses H, evidence e, and context c of interest to an analyst – so that the AI can usefully assist by computing numerical likelihoods of relevance to the analyst.

Other opportunities for cognitive assistance may exist at the HELP steps corresponding to hypothesis generation, evidence isolation, and posterior aggregation. But hypothesis generation and evidence isolation appear especially difficult to automate, because they depend on the vast human knowledge of categorical H, e, and c entities noted above. Of all the steps, posterior aggregation appears to be most amenable to automation, for two reasons. First, posterior aggregation is merely a calculation performed using the products of the previous three steps, so it is easy to automate. Second, the calculation is difficult for unaided humans, as evidenced by the heuristics and biases measured in numerous experiments.

The most common bias in posterior aggregation is *conservatism* (Edwards 1982; Edwards et al. 1968), where the human heuristic is an arithmetic average of prior and likelihood – rather than a normalized product of prior and likelihood. The upshot is that humans fail to extract all the certainty that is available from likelihoods in sequential updating, so they seek more evidence than is needed to achieve a requisite level of certainty. An automated system to sup-

port posterior aggregation, in the fourth step of HELP, would allow the human-system together to more efficiently manage the collection and exploitation of information in multi-source intelligence missions.

One such system, using visualizations of Bayesian computations to intuitively illustrate posterior aggregation, has been implemented as a prototype dubbed Bayesian Boxes (Burns 2006). The same system has been validated in human experiments (Burns 2007), which demonstrated that Bayesian Boxes are effective in mitigating a number of cognitive heuristics and biases, including anchoring and conservatism.

Notice that the above ideas for cognitive assistance differ from existing AI systems that employ Bayesian Networks (Jensen 1996). These systems are useful for Bayesian propagation of probabilities through large networks of nodes (representing hypotheses and evidence) and arcs (representing likelihoods). But such systems require that the network structure and input values be supplied by users (Karvetski et al. 2013), and in that sense humans are assisting the machine, rather than the other way around – at least in the three HELP steps of hypothesis generation, evidence isolation, and likelihood estimation. The users who develop the networks and inputs are usually Bayesian experts themselves, which is why they have chosen to employ Bayesian Networks in the first place. Therefore these systems are not benefiting the vast majority of analysts who are most in need of cognitive assistance.

Dozens of interviews with all-source analysts (Burns 2014b) indicated that their inferences involve only a few hypotheses and a few items of evidence – similar to the case study of *Ominous Airplanes*. The interviews also illustrated the importance of abduction, whereby an analyst generated a pivotal hypothesis that was not considered in an earlier mental model – similar to "It's Students" in *Ominous Airplanes*. For these prototypical problems of all-source intelligence, Bayesian Networks offer no advantage, because there is no need to perform posterior aggregation over large networks of pre-defined hypotheses and evidence. Instead analysts could benefit from tools for structuring and supporting their own abductive inferences, as illustrated by HELP in the example of *Ominous Airplanes*.

References

Bayes, T. 1763. An Essay Toward Solving a Problem in the Doctrine of Chances. *Philosophical Transactions* 53:370-418.

Beebe, S. and Pherson, R. 2012. *Cases in Intelligence Analysis: Structured Analytic Techniques in Action*. Los Angeles: Sage CQ Press.

Burns, K. 2015. Computing the Creativeness of Amusing Advertisements: A Bayesian Model of Burma-Shave's Muse. *Artificial Intelligence for Engineering Design, Analysis, and Manufacturing* 29:109-128.

Burns, K. 2014a. Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking (ICArUS): A Computational Basis for ICArUS Challenge Problem Design. MITRE Technical Report, MTR 149415, McLean, VA.

Burns, K. 2014b. Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking (ICArUS): Phase 2 Challenge Problem Design and Test Specification. MITRE Technical Report, MTR 149412, McLean, VA.

Burns, K. 2007. Dealing with Probabilities: On Improving Inferences with Bayesian Boxes. In Hoffman, R., ed. *Expertise Out of Context*, 263-280. New York: Lawrence Erlbaum.

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, R., and Brehmer, B., eds. *How Professionals Make Decisions*, 15-28. Mahwah, NJ: Lawrence Erlbaum.

Edwards, W. 1982. Conservatism in Human Information Processing. In Kahneman, D., Slovic, P., and Tversky, A., eds. *Judgment Under Uncertainty: Heuristics and Biases*, 359-369. Cambridge: Cambridge University Press.

Edwards, W., Phillips, L., Hayes, W., and Goodman, B. 1968. Probabilistic Information Processing Systems: Design and Evaluation. *IEEE Transactions on Systems, Man, and Cybernetics* 4:248-265.

Fischhoff, B. and Beyth-Marom, R. 1983. Hypothesis Evaluation from a Bayesian Perspective. *Psychological Review* 90:239-260.

Heuer, R. 1999. *Psychology of Intelligence Analysis*. Center for the Study of Intelligence, Central Intelligence Agency, Washington, DC.

IARPA. 2010. Integrated Cognitive-neuroscience Architectures for Understanding Sensemaking (ICArUS), Intelligence Advanced Research Projects Activity (IARPA), Broad Agency Announcement (BAA), IARPA-BAA-10-04.

Jensen, F. 1996. *An Introduction to Bayesian Networks*. New York: Springer

Karvetski, C., Olson, K., Gantz, D., and Cross, G. 2013. Structuring and Analyzing Competing Hypotheses with Bayesian Networks for Intelligence Analysis. *EURO Journal on Decision Processes* 1:205-231.

Klayman, J. and Ha, Y. 1987. Confirmation, Disconfirmation, and Information in Hypothesis Testing. *Psychological Review* 94: 211-228.

Klein, G., Phillips, J., Rall, E., and Peluso, D. 2007. A Data-Frame Theory of Sensemaking. In Hoffman, R., ed. *Expertise Out of Context*, 113-155. New York: Lawrence Erlbaum.

Klein, G., Moon, B., and Hoffman, R. 2006a. Making Sense of Sensemaking 1: Alternative Perspectives. *IEEE Intelligent Systems* 21:70-73.

Klein, G., Moon, B., and Hoffman, R. 2006b. Making Sense of Sensemaking 2: A Macrocognitive Model. *IEEE Intelligent Systems* 21:88-92.

Mueller, S. 2009. A Bayesian Recognitional Decision Model. *Journal of Cognitive Engineering and Decision Making* 3:111-130.

Nickerson, R. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2:175-200.

Pirolli, P. 2007. *Information Foraging Theory: Adaptive Interaction with Information*. New York: Oxford University Press.