Confirmation Bias in the Analysis of Remote Sensing Data

Paul E. Lehner, Leonard Adelman,
Robert J. DiStasio Jr., Marie C. Erie, Janet S. Mittel, Sherry L. Olson

Abstract

In practical application, analysis of remote sensing data requires a mix of technical analysis and best expert judgment. Unfortunately, a substantial experimental literature on judgment indicates that expert judgment is systematically flawed. In particular, experts are prone to a confirmation bias – where focus on a proposed hypothesis leads the expert to seek and overweigh confirming versus disconfirming evidence. In remote sensing, this predicts a tendency toward false positives in interpretation - concluding the evidence supports a hypothesis when it doesn't. In this paper, we empirically examine confirmation bias in technical data analysis, along with an approach to mitigating this bias that systematically promotes consideration of alternative causes in the analysis. Results suggest that analysts do exhibit confirmation bias in their technical analysis of remote sensing data; and furthermore that structured consideration of alternative causes mitigates this bias.

Introduction

Consider the following hypothetical scenario. As a result of recent flooding, there are hundreds of possible locations where caustic chemicals may be leaking into the waterways. Scarce containment resources must be quickly dispatched to the few locations where spills have actually occurred. Response time is critical. At the site of a particular chemical plant there is concern that ethylene glycol, a dangerous substance, is leaching into the waterways upstream from a residential area. A flyover was conducted in response to the flooding, where a small plume of discolored water was visible from the air. Spectral data from this flight was processed to examine the area of the observed plume. Analysis of the spectral sensor data is a highly *technical data analysis* task requiring a specialist in spectral remote sensing, sophisticated algorithmic processing of the data, and expert determination of whether the data matches the spectral signature (acquired from laboratory experiments) of ethylene glycol. A technical analyst reviews the data, concludes there is a match, and asserts that she is 85% confident that ethylene glycol is leaking at the site.

The conclusion "85% confident" appears to be the result of algorithmic and deductive analysis. In fact, it is the product of highly subjective judgments on the part of the analyst. To begin with, the analyst has no way of confirming that the spectral sensor was in proper working condition. She can only take that on faith. Second, there are other benign substances that have a similar spectral signature. The analyst also does not know

if any of these substances are produced at the plant. Furthermore, there are numerous other substances in and around the plant that are likely to have leached into the water. She is not sure what chemicals are there. She cannot be sure that the library of spectral signatures in her possession contains all of these substances, so it's possible that some of the spectral readings are from other chemicals. The analyst understands these uncertainties, accounts for them in her "best expert judgment" and asserts "85% confident".

This scenario is characteristic of many technical data analysis problems: time is precious, data is scarce, the environment cluttered, decision makers are pressing for an answer and the analysts must rely on their own subjective expert judgment to generate needed answers.

To our knowledge, subjective judgment in the analysis of remote sensing data has not been systematically examined. This is unfortunate, since subjective expert judgment is systematically flawed.

There is a considerable research literature on human judgment, including expert judgment, which shows that judgments consistently exhibit well-known biases (Bazerman, 1996; Hastie & Dawes, 2001; Nickerson, 1998). A bias of particular concern to analysis is the *confirmation bias*, where once someone begins to focus on a single hypothesis, there is a natural tendency to seek and overweigh confirming rather than disconfirming evidence. Research with intelligence analysts confirms that confirmation bias (Tolcott, et. al. 1989, Lehner, et. al., 2006) occurs in complex analysis tasks. In the above scenario, for example, just the fact that the analyst was asked to consider the "ethylene glycol" hypothesis may be enough to bias the analyst toward confirming that hypothesis.

Unfortunately, reducing confirmation bias is difficult. Making people aware of it, and instructing them to try to be unbiased, does not mitigate it (Burke, 2006; Nickerson, 1998). There is however some research that suggests that requiring people to consider *alternative hypotheses* or scenarios does mitigate the confirmation bias somewhat (Galinsky and Moskowitz 2000; Evans et al., 2002; Lehner et al., 2006). In this *alternative hypotheses* approach, people are required to enumerate one or more alternatives to their favored hypothesis and then to systematically evaluate each evidence item against each alternative hypothesis. One instantiation of this approach, called Analysis of Competing Hypotheses (ACH), is gaining popularity in the intelligence community (Heuer, 1998; Heuer, et.al., 2004; Jones 1999).

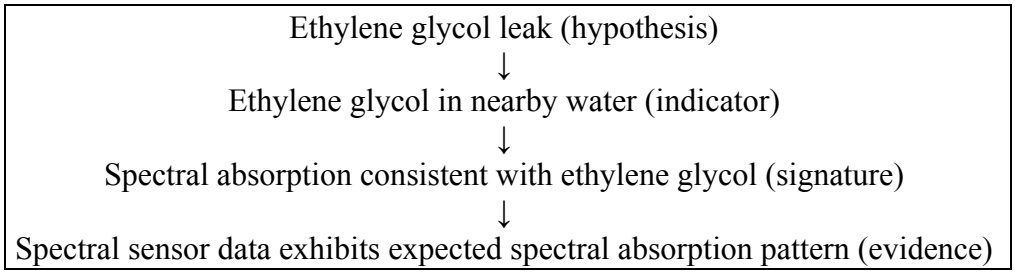
There are reasons however to question the general effectiveness of the *alternative hypotheses* approach, at least as it is applies to sensor analysis. First, the impact on bias reduction is not robust. In the Lehner experiment, for example, ACH only reduced bias in non-experts. Second, anecdotally one hears of many instances where analysts simply "go through the motion" of proposing and rejecting alternative hypotheses. They have a favored hypothesis and they view a comparable analysis of alternative hypotheses as a waste of precious time and effort. Third, alternative hypotheses are usually at the same level as the original hypothesis (leak vs. no leak), whereas sensor analysts consider a

diversity of local reasons why the sensor return may be faulty (atmospheric effects, other chemicals, etc.). The sensor analysts task is not to determine if the hypothesis is true, but to provide an accurate interpretation of the sensor data.

In this paper we propose and test an alternative to the *alternative hypotheses* approach, called the *alternative causes* approach, that we believe is appropriate for technical data analysis of sensor returns. In the *alternative causes* approach analysts are asked to focus on their favored hypothesis. They are asked to document the causal sequence they wish to claim and are then asked to enumerate alternative causes that could "explain away" different elements of their proposed causal sequence.

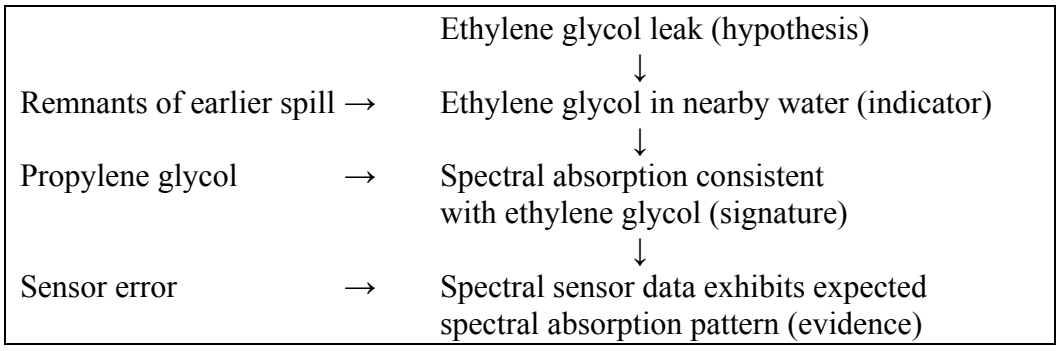
To illustrate the *alternative causes* approach, consider again the example in the introduction. An analyst has been asked to determine if ethylene glycol is leaching into the waterways at a chemical plant. A spectral analyst has received sensor readings that suggest greater spectral absorption in the bands normally associated with ethylene glycol. The causal chain she proposes is shown in Figure 1.

Figure 1: Causal Chain from Hypothesis to Evidence



In short, the sensor readings were caused by the hypothesis and therefore are evidence that the hypothesis is true. The analyst is then asked to also document possible alternative causes of each step in the suggested causal chain. In this case the alternatives may be as shown in Figure 2.

Figure 2: Alternative Causes for Positive Evidence
(Positive Sensor Analysis Template)



As can be seen, rather than articulate and evaluate an alternative to the core hypothesis (Ethylene glycol leak), experts are asked to enumerate possible alternative causes that could explain away each step in the causal reasoning chain. Logically, if any one of these alternative causes is true, then the observed sensor reading provides little, if any, evidential support for the hypothesis.[1]

In the experiment discussed below, we implemented the alternative causes approach by having analysts document their reasoning using Sensor Analysis Templates that follow the structure shown in Figure 2. The example above is an instance of a positive template, where the evidence supports the hypothesis. A similar structure is used for negative evidence, as illustrated in Figure 3. In the negative template the proposed causal chain down the right side of the template is the same as the positive template except for the negative evidence at the bottom right. The alternative causes now provide possible explanations for why components in the intended causal chain may be true **despite** the negative evidence. In this case, the failure to detect the spectral absorption pattern for ethylene glycol may be due to a sensor failure. Alternatively, the flooding may have diluted the concentration to the point where an airborne sensor cannot detect it.

Figure 3. Alternative Causes for Negative Evidence
(Negative Sensor Analysis Template)

		Ethylene glycol leak (hypothesis)
		↓
No plausible alternative	→	Ethylene glycol in nearby water (indicator)
		↓
Flooding dilutes chemicals	→	Spectral absorption consistent with ethylene glycol (signature)
		↓
Sensor error	→	Spectral sensor data does **not** exhibit expected spectral absorption pattern (evidence)

Given past research on confirmation bias, it can be argued that the *alternative causes* approach, because it encourages analysts to focus on a single hypothesis, would increase rather than mitigate confirmation bias. Our research hypothesis is rather the opposite. Specifically we believe that encouraging analysts to consider alternative causes in their analyses will often lead them to deduce that the causal link between their favored hypothesis and the supporting evidence is weak. They will therefore be less subject to "false positives," where they incorrectly interpret data as favoring a hypothesis when it doesn't.

---

[1] The structure mimics the elements likely to be found in a Bayesian network (Neapolitan, 2003) representation of a sensor interpretation problem. In fact, this structure was initially designed to encourage analysts to reason through the causal considerations in a Bayesian network, without requiring the pre-specification of alternatives and the numerous numerical judgments needed to populate a Bayesian network.

The experiment reported below tests three research hypotheses:

1. Sensor analysts are naturally biased toward interpreting sensor data as confirming a hypothesis. In cases where the hypothesis is wrong, this will lead to false positives.

2. Requiring analysts to enumerate alternative causes will increase the number of alternative causes they consider.

3. Requiring analysts to enumerate alternative causes will decrease the number of false positives.

Method

Ten participants with professional experience in spectral analysis analyzed six test problems. Five participants used the Sensor Analysis Templates to document their reasoning, five did not. Provided below is a detailed description of the experimental method. This description is divided into the following subsections: design, participants, problems, conditions, and procedures.

Design: Participants were first matched for overall experience in performing sensor analysis, and then randomly assigned to either the Template or No Template condition. The Template participants stepped through the process of enumerating alternative causes. All participants worked all six analysis problems. For each problem, the analyst was given a question/hypothesis and asked to determine if the evidence supported or contradicted that hypothesis. For half the problems, the evidence provided supported the hypothesis (Normatively True) and for half the problems the evidence contradicted the hypothesis (Normatively False). So, this is a 2x2 design where Template (Yes or No) is a between-subject variable and Hypothesis (Normatively True or Normatively False), is a within-subject variable. Saying that the Normatively False hypothesis is true is a false positive.

Participants: Ten experienced spectral analysts at The MITRE Corporation participated in the experiment. None of the participants were familiar with the *alternative causes* approach or the six specific problems used in the experiment. However, all participants were familiar with the spectral sensors described in the six problems, so they were comfortable addressing the sensor interpretation tasks presented in the problems.

Participants did, however, vary in their overall level of experience. Consequently, we first divided the ten participants into three groups: those with less than three years of experience (n = 2), those with three to ten years of experience (n = 6), and those with more than ten years of experience (n = 2). Then, participants in each of these groups were randomly assigned to the Template and No Template conditions, resulting in five participants in each condition. The order in which the participants participated in the experiment was determined by their availability.

Problems: Two spectral analysts on the research team developed six analysis problems. The problems were designed so that the hypothesis was supported for three of the problems (called Normatively True problems) and contradicted for the other three problems (called Normatively False problems). The participating analysts had to conclude if the stated, problem hypothesis was true or false based on provided information for each problem. The name and hypothesis for each of the problems is listed below

Normatively True problems:

- Stressed Vegetation - There is a region of stressed vegetation at the location of a river inlet
- Cameron, LA – There is a Malathion leak
- Another Pipeline Leak - The stain on the ground is a new spill.

Normatively False problems:

- Port Arthur - A petrochemical plant is polluting the coastal area with Ethylene Glycol
- Oil Containment Breach - There are hydrocarbons on the water outside the protective berm surrounding an oil tank
- Mineral Mapping - Jarosite and goethite are present in the scene

In all cases a problem was presented with accompanying background information. For example, in the Another Pipeline Leak problem, participants were told that a pipeline transshipment point facility has had a history of pipeline leaks and there are now reports of a new leak. Then, participants were given the hypothesis. For the Another Pipeline Leak problem, the hypothesis was that an area of stained ground was a new oil spill. Participants were not asked to perform any algorithmic analysis of their own but, instead, were asked to interpret the results of various algorithmic analyses to reach their conclusion. For the Another Pipeline Leak problem, for example, they were given COMPASS imagery and in-scene and library spectral data for the soil-vegetation ground cover, the soil, the tainted areas, a nearby pond, the pipeline and trees, and library spectral data of the vegetation, hydrocarbons, and asphalt.

We did not want ceiling effects, where all problems could be solved correctly, or floor effects, where none of them could be solved. So, the spectral analysts on the research team developed what they considered to be difficult, but reasonable problems to solve given that the participating analysts could not do any further algorithmic analysis on the provided data. Pilot-testing lead to minor revisions to the wording for some of the problems so that both spectral analysts on the research team thought the problems met our criteria.

The order in which the ten participants received the six problems was randomized in an effort to ensure that no problem was always the first or last one solved by all participants.

This was done to minimize the possibility that results for specific problems were due to fatigue, learning effects (with the template), or any other factor not explicitly controlled for in the experiment.

Template and No Template Conditions: Based on the provided data for each of the six problems, participants in both conditions had to (1) determine whether they thought the hypothesis was true or false (i.e., they were forced to make a decision), (2) indicate the "amount of support" they thought the evidence provided either for or against the hypothesis and (3) type the rationale for their conclusion. "Amount of support" was determined by participant's answer on the following seven-point scale: Very Strong Support *For* Hypothesis, Strong Support *For* Hypothesis, Some Support *For* Hypothesis, Equal Support *For and Against* Hypothesis, Some Support *Against* Hypothesis, Strong Support *Against* Hypothesis, and Very Strong Support *Against* Hypothesis.

Participants in the No Template condition could use whatever mental hypothesis-testing processes they wanted to interpret the presented data and solve the six problems. However, participants in the Template condition had to complete positive and negative evidence templates, respectively, for each of the six problems. Template participants were first given a tutorial explaining how to complete the templates, and two simple problems to get practice using them, before using the templates to help solve the six experimental problems. Participants in the No Template group also worked the two practice problems.

Procedurally, the goal of the template is first to decompose the hypothesis-testing process into its component parts and then, to generate alternatives for each part. So, with decomposition in mind, participants were taught to complete the templates by first going down the right-hand side of the template. The hypothesis block was specified for all Template (and No Template) participants to make sure that everyone started with the same, exactly-worded hypothesis for the problem. Then, Template participants were asked to specify an indicator for that hypothesis, the signature features that they would expect to see if the indicator was true and, then, the positive evidence in the problem they actually saw as supporting the signature and, by inference, the indicator and hypothesis.

Then, Template participants moved up the left-hand side of the template trying to generate alternatives causes to explain the positive evidence (i.e., alternative causes or reasons why one could have found the positive evidence even if the signature was actually false), the signature (i.e., alternative causes or reasons why signature may be true even if the indicator was false) and lastly, the indicator (i.e., alternative causes or reasons the indicator could be true even if the hypothesis was false). If participants could not think of an alternative cause, they were instructed to type "No Alternative" in the appropriate box. After listing alternative causes in a particular box, participants indicated how likely they thought the alternative causes were using a five-point scale (very likely, somewhat likely, 50/50, somewhat unlikely, and very unlikely), and how informed they considered themselves to be to make these judgments based on their experience and the provided information using a three-point scale (expert and ample data, expert or ample data, and neither expert nor have ample data).

Then, the Template participants used the same basic procedures to complete the negative evidence template. The indicator and signature entries in the negative evidence template were the same as in the positive evidence templates, so Template participants typically copied and pasted their entries from the positive to negative evidence templates. When completing the negative evidence box, participants were instructed to list evidence that they did not see but expected to see if the hypothesis was true, as well as evidence they did see that did not support the hypothesis. Then, moving up the left-hand side of the negative evidence template, participants listed (negative) evidence alternatives (i.e., alternative causes or reasons for how one could obtain the negative evidence even if the indicator and, by inference, the hypothesis was true), signature alternatives (i.e., alternative causes for how the signature could be false even if the indicator was true), and indicator alternatives (i.e., alternative causes for how the indicator could be false even if the hypothesis was true). Again, participants could say "No Alternatives" if they could not think of any for an alternative box. And as with the positive template, after listing alternative causes in a particular box, participants indicated how likely they thought the alternative causes were, and how informed they considered themselves to be in making the judgments.

Procedures: Participants in the Template and No Template conditions followed the same procedures, except for those that were unique to the Template condition. Procedurally, each session had the following parts:

- Background where the session was described and participants signed the informed consent form;
- Template tutorial and illustrative example problem for Template participants or just the illustrative, example problems for the No Template participants;
- Two practice problems, which were the same for both groups, but Template participants used the templates to work them;
- The six problems, either worked with or without template based on condition; and
- Questionnaire and concluding interview

All participants worked the two practice and six experiment problems individually in a private room. The practice problems were much easier than the six actual problems. This was particularly important to ensure that the Template participants knew how to use the templates. However, the practice problems also ensured that the No Template participants knew what was expected of them and that they had the same amount of practice as the Template participants. Template participants were also shown how the spectral analysts on the research team had completed the templates for the practice problems to help them understand how to complete the templates.

Problem information was presented and responses recorded using slides in a PowerPoint file for each problem and coded separately for each participant. A video recorder (with microphone) was used to record all comments each participant made and the slides they

were looking at when they made their comments. In addition, paper copies of the presented problem information (and template tutorial for the Template participants) were available if participants wanted to look at them.

All participants were asked to think aloud when they worked each problem. Thinking aloud was necessary for the No Template condition in order to identify the different alternatives that these participants naturally thought of while working each problem. Since it was possible that thinking aloud could have affected the participating analysts' hypothesis-testing process, the Template participants also were required to think aloud as they worked each problem to make the two conditions comparable. The only procedural difference that we wanted between the two conditions was whether or not the participant used the templates. Slides were repeatedly embedded in the file to remind the participants to think aloud. In addition, a member of the research team was present in the experimental room to remind the participants to think aloud if they forgot, and to answer any procedural questions or provide procedural support. Except for one case, the research team member in the room was not a spectral analyst. This was done to minimize the chance that the team member would inadvertently provide substantive support to the participants.

Participants were asked to answer a questionnaire after completing all six problems. The questionnaire for the Template participants had an initial set of ten questions requesting their opinion of the templates' value in doing sensor analysis, thinking of alternative causes, time commitment, ease of learning, communication value, organizational fit, etc. These questions required answers on a 5-point (Likert) scale going from strongly agree to strongly disagree, so lower numbers indicated more favorable opinions. Participants in both conditions answered questions about the adequacy of the six problems, their experience with the topic areas covered in each of the problems, and their education and professional experience.

The session concluded with a brief interview session with one of the research teams' spectral analysts to obtain any additional comments and insights about the session, and to answer any remaining questions. All participants were requested not to talk with their peers about the experiment in an effort to ensure that other participants' interpretation processes were unaffected when they began the session. Four hours were scheduled for each participant, but each participant could take as much time as he or she wanted to complete the entire session. Breaks were scheduled every hour and a half, but were taken by each participant at their discretion.[2]

---

[2] We note here that we tried-out or "pilot-tested" both conditions with one additional participant in each condition. There were only two pilot-test participants because of the limited availability of spectral analysts. Nevertheless, pilot testing was extremely valuable and led to subsequent improvements in implementing both conditions as well as wording the six problems to minimize any misunderstandings.

Results

The analysis below focuses on the following three experimental hypotheses, which correspond to the research hypotheses listed in the introduction:

> H1: Participants in the No Template group are biased toward interpreting sensor data as confirming a hypothesis, resulting in false positives for the Normative False problems
>
> H2: Participants using the Template will consider more alternative causes than participants in the No Template group.
>
> H3: Requiring analysts to enumerate (more) alternative causes will result in participants using the Template to generate fewer false positives on Normative False problems than participants in the No Template group.

Alternative Causes Considered

The intent of the template is to encourage analyst to consider alternative causal explanations for the data observed. As noted above, participants were video taped and were repeatedly asked to talk out loud as they worked through each target folder. Everything verbalized by the participants was transcribed. Each statement in the transcription was coded as a statement about an alternative cause, an evidence or explanatory statement, or other. All written statements, either as part of the rationale or template, were coded in the same manner.

Coding was done by the two team members with professional experience in hyperspectral analysis. Several target folders were independently coded by both team members. Inter rater reliability, for all codes, was .75 or higher.

After initial coding, all statements of alternative causes were carefully examined to remove duplicates. Many duplicates were the result of participants saying word for word exactly what they were writing as they filled out the templates and rationale. Others were the results of participants thinking about the same alternative cause at different times while solving the problem.

Table 1 shows the average number of unique alternative causes expressed by participants either verbally, in their written rationales, or in the templates. The data is partitioned by whether or not the participants used the Template and whether the data in the problem supported the suggested hypothesis or contradicted it (Normative True vs. Normative False). A mixed Analysis of Variance (ANOVA) shows that Template participants generated more alternatives than No Template participants ($p=.010$, 1-tailed) and that there were no other main or interaction effects.[3] Note that Template participants

[3] Template vs. No Template is a between participants factor. Normative True vs. Normative False is a within participant factor.

generated more alternatives than No Template participants for each of the six target folders and four of the six were found to be, by themselves, statistically significant.

Table 1. Average number of alternative causes per problem expressed by participants

	Normative True	**Normative False**
Template	6.2	7.3
No Template	3.1	2.7

The difference between the number of alternative causes expressed by Template and No Template participants appears attributable to the number of additional causes written in the templates. On average there were 4.87 alternative causes written in the templates for each analysis problem.

Overall the data support the experiment hypothesis (H2) that participants using the template considered more alternatives causes.

Performance

Research hypothesis H1 is that No Template participants will be prone to incorrectly interpreting the data in the Normative False problems as supporting the hypothesis; that is they will be prone to false positives. Research hypothesis H3 is that template users will generate fewer errors, specifically fewer false positives, than No Template users.

For each problem, participants were asked to decide if the suggested hypothesis is True or False. Table 2 shows the number of correct and incorrect answers partitioned by Group and Normative Answer. As can be seen, when the evidence supported the suggested hypothesis (Normative True) the participants in both the Template and No Template group usually inferred the correct answer. However, as predicted, when the evidence contradicted the suggested hypothesis (Normative False) the Template group appears more likely to solve the problem correctly than the No Template group. (Statistical significance is discussed below.)

Table 2. Frequency of correct and incorrect conclusions

	Normative True			Normative False	
	Conclude "True"	Conclude "False"		Conclude "True"	Conclude "False"
Template	13	2		2	13
No Template	12	3		7	8

Note that all participants were required to make a final "True" vs. "False" judgment for each problem. However in several instances participants felt that the evidence, on balance, neither supported nor contradicted the suggested hypothesis. In these instances, their selection of "True" vs. "False" was arbitrary. A better measure of the participants' interpretation of the evidence is whether they entered a positive or a negative value for support.

Table 3. Frequency of correct and incorrect interpretations

	Normative True			Normative False		
	"Positive Support"	Neither Pos or Neg	"Negative Support"	"Positive Support"	Neither Pos or Neg	"Negative Support"
Template	12	2	1	1	1	13
No Template	12	2	1	6	3	6

Table 3 shows the frequency of positive, zero and negative support values partitioned by Group and Normative Answer.[4] Here again, when the suggested hypothesis is Normative False, the Template participants were more likely to conclude that the evidence supported another conclusion than the No Template group.

Although Tables 2 and 3 are visually compelling, direct statistical analysis of these contingency tables is inappropriate since each participant solved all six problems. Instead we performed a mixed ANOVA analysis. Table 4 shows the average number of support judgments in the correct direction by condition. No Template participants under the Normative False condition interpreted the data correctly only 40% of the time.

Table 4: Average number of correct interpretations

	Normative True	Normative False
Template	2.4	2.6
No Template	2.4	1.2

The mixed ANOVA analysis showed a significant effect for the template ($p=.036$, 1-tailed) and an interaction effect ($p=.022$, 1-tailed). No Template participants did significantly better on the Normative True problems than on Normative False problems ($p=.035$, 1-tailed, paired t-test). Template participants did significantly better than No

[4] Support scores varied on a seven point scale, ranging from very strong support for a hypothesis to very strong support against the hypothesis. However, in this analysis we only examined the direction of support (+ vs. -) and ignored the degree of support. This is because we did not have normative criteria for degree of support. Although a priori we knew, for each folder, whether the evidence supported or contradicted the suggested hypothesis (Normative True vs. Normative False) we could not assert the appropriate degree of support. A determination of "some support" score for a Normative True problem is considered as valid as a "very strong support" score. They were counted the same.

Template participants on the Normative False problems (p=.017, 1-tailed, independent t-test).

We note also that the template participants did better than the No Template on all three Normative False problems. There was one problem, however, where the difference was especially pronounced. As shown in Table 5, for the Oil Breach problem, none of the No Template participants assigned negative support while none of the Template participants assigned positive support (p=.048, 2x3 Fisher exact[5], 2-tailed).

Table 5: Frequency of positive and negative support values for Oil Breach problem. (Correct answer is "Negative Support")

	"Positive Support"	Neither Pos or Neg	"Negative Support"
Template	0	1	4
No Template	3	2	0

Overall, the results are consistent with the prediction (H1) that analysts are naturally biased toward false positives. The results also support experimental hypothesis H3 that use of the template significantly reduces this bias.

Additional analyses

The results presented above examined the three experiment hypotheses. Below we examine some ad hoc results.

Time to Completion.

Participants in the No Template group took on average 19 minutes to complete each problem. Participants in the Template group took an average of 31 minutes. A mixed ANOVA analysis showed a main effect for Template vs. No Template (p<.02, 2-tailed), and no other effects.

"Errors" in completing the templates

Although the templates were obviously usable, they were not easy for participants to fill out. Participants made a number of errors in completing the templates. Logical errors occurred when participants wrote something that made no sense, such as asserting that X may be an alternative cause of Y instead of X. Approximately 15% of the statements

[5] Fisher exact test are used to statistically analyze contingency tables with small samples sizes, such as Table 5.

written in the alternative causes boxes were counted as logical errors. These were not included in the analysis above.

Participants also made placement errors where alternative causes made sense, but were entered into the wrong box. About 16% of the meaningful alternatives entered into the templates were placement errors.

Impact of Experience

Participants self-rated their years of experience in three separate areas: Imagery Analysis, Remote Sensing Analysis and domain-specific analysis experience in each of the six problem domains.

All subjects had one or more years of experience in remote sensing and in each of the six problem domains. Nine of the ten participants had at least one year of experience in imagery analysis. In short, all participants had some professional experience in the relevant areas. Table 6 shows the average years of experience by each category.

Table 6: Average years of experience by Group

	Imagery Analysis	Remote-sensing Analysis	Average Domain Experience
Template	6.8	6.4	1.9
No Template	4.8	6.8	2.5

Table 7 shows the correlation between each type of experience and the support values partitioned by group and Normative answer. If experience improves performance, then for the Normative True problems there should be a positive correlation between support values and experience. For Normative False problems there should be a negative correlation.

Table 7: Correlations between support values and experience

	Template		No Template	
	Normative True	Normative False	Normative True	Normative False
Imagery	0.338	0.066	-0.116	0.014
Remote	0.048	0.505	-0.277	0.103
Domain	-0.249	0.274	0.086	-0.036

None of the correlations were significant in the expected direction. The only correlation that approached significance (0.505) was in the wrong direction.

Overall we found no consistent relationship between experience and performance on the analysis problems. This is not too surprising since all of the participants had significant professional analysis experience in relevant areas. Some were more experienced than others, but none of the participants were novices.[6]

Analysis of Questionnaire.

At the end of the experiment, Template participants where asked to complete a questionnaire about their view of the template. Overall participants viewed the Template positively. The average response of all questions was below 3 except for Question 5 – "It was easy to learn how to use the templates" which had an average score of three.

Of interest is the relationship between participant experience and how a participant viewed the Template. In general one might expect that more experienced analysts would find the Template less useful than less experienced analysts. They are more set in their ways.

Table 7 shows the Imagery and Remote Sensing experience of the 5 Template participants and their average score on the ten questions. As expected there is a positive correlation between Imagery and Remote Sensing Analysis experience and the response to the questionnaire, .53 (not significant) and .72 (p=.042, 1-tailed) respectively.

Table 7 Experience and Response to Questionnaire

Imagery Analysis Experience	Remote Sensing Analysis Experience	Average response to questionnaire (1 = best, 5 = worst)
0	3	1.8
1	1	2.3
10	15	2.6
20	10	2.3
3	3	1.8

Discussion

There are three principal results. First, sensor analysts were naturally prone to falsely interpret sensor data as confirming a hypothesis. In this experiment, just the fact that they were asked to confirm or disconfirm a suggested hypothesis made them prone to misinterpret the evidence as confirming the suggested hypothesis. This represents a

[6] All of the target folders involved hyperspectral imagery (HSI) analysis, yet we failed to ask specifically about HSI experience. As an ad hoc analysis we obtained ratings on the level of HSI experience for the 10 participants and again found no evidence of improved performance with increased experience. Note that all of the participants had some professional experience in HSI analysis.

confirmation bias, found in many other expert judgment tasks (Lehner et al., 2006; Nickerson, 1998). Second, use of the Template led analysts to consider more alternative causes in their reasoning. Third, use of the Template significantly reduced the instances where analysts falsely interpreted sensor data as confirming a suggested hypothesis. The benefits of the template were obtained even though participants had difficulty filling out the templates and made a number of mistakes when doing so.

The fact that use of the templates improved performance, even though the templates were often filled out incorrectly, suggests that the benefits were obtained from the reasoning process invoked by the templates. Participants using the templates did better because they devoted more time and effort to reasoning about alternative causes. The templates helped analysts regardless of their experience level, even though more experienced analysts thought that the templates were less useful than less experienced analysts.

Regarding future work, these results need replication. Our research was focused on developing and testing methods to support professional analysts engaged in technical analysis of sensor returns. This narrow focus naturally limited the population of possible participants. However, there is nothing in the approach we developed that is specific to just professional sensor analysts. Studies could be performed examining the effectiveness of the alternative causes approach in a larger population of more traditional participants (e.g., students) and with other areas of analysis (e.g., all-source analysis).

There are a couple of issues that merit further investigation. First, a colleague of ours who reviewed this work noted that the hypotheses we asked participants to evaluate reflect circumstances where serious negative consequences would result if the hypothesis were missed. Failing to detect a chemical spill, for example, could result in lives harmed. By contrast, a false alarm on a chemical spill would have relatively minor consequences (e.g., wasted resources in checking out the spill). Consequently, it could be argued that the tendency of the No Template participants to "misinterpret" evidence as confirming a hypothesis was not an error at all, they were just being prudently cautious. This explanation doesn't apply to all of the problems, nor does it account for the improved performance of the Template participants, but it does raise the question of whether the pattern of bias would differ if the consequences for error differed. If so, would use of the template result in better judgments if there was a bias toward disconfirming hypotheses?

Previous research on confirmation bias often begins with hypotheses that participants initially generated or selected. By contrast, in this experiment participants were given the hypotheses to evaluate. The hypotheses were not in any way "preferred" or "favored" by the participants. As is typically found, confirmation bias manifested anyway. However, this does bring up the question of whether or not the debiasing effect of the *alternative causes* approach generalizes to circumstances where participants are assessing their preferred hypotheses.

Assuming these results are robust and generalize, the next concern is simplification. The current method, although useable, is still awkward for analysts to use. As we've noted repeatedly, the value of the method appears to derive from the fact that it encourages

analysts to think through alternative explanations of the evidence. Surely there are simpler ways to achieve this behavior, while still obtaining the debiasing benefits. Simplification will increase the chances that such debiasing methods will become part of routine analytic practice.

References

Adelman, L., Gualtieri, J., & Stanford, S. (1995). Examining the effect of causal focus on the option generation process: An experiment using protocol analysis. *Organizational Behavior and Human Decision Processes, 61*, 54-66.

Bazerman, M.H. (2006) *Judgment in Managerial Decision Making* (6th Ed.). NY: Wiley.

Burke, A.S. (2006), Improving prosecutorial decision making: some lessons of cognitive science. *William and Mary Law Review*, vol. 47, pp. 1587-1634.

Evans, J. St. B.T., Venn, S., and Feeney, A. (2002), "Implicit and explicit processes in a hypothesis testing task," *British Journal of Psychology*, vol. 93, pp. 31-46.

Galinsky, A. D., and Moskowitz, G.B. (2000), "Counterfactuals as behavioral primes: Priming the simulation heuristic and consideration of alternatives," Journal of *Experimental Social Psychology,* vol. 36, pp. 384-409.

Hastie, R., & Dawes, R.M. (2001). *Rational Choice in an Uncertain World*. Thousand Oaks, CA.

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

Heuer, J.R., Good, L., Shrager, J., Stefik, M., Pirolli, P., and Card, S., *ACH0: A Tool for Analyzing Competing Hypotheses,* (Draft Technical Report). Palo Alto, CA: PARC, 2004.

Jones, M.D. *The Thinker's Toolkit.* NY: Three Rivers Press, 1998.

Koehler, D.J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110*, 499-519

Lehner, P., Adelman, L, Cheikes, B. and Brown, M. (2006) Confirmation bias in complex analyses, submitted *IEEE Transactions on Systems, Man and Cybernetics*.

Mehle, T. (1982). Hypothesis generation in an automobile malfunction inference task. *Acta Psychological, 52*, 87-106.

Neapolitan, R.E. *Learning Bayesian Networks*. Prentice Hall, 2003.

Nickerson, R.S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175-220.

Tolcott, M.A., Marvin, F.F.,, and Lehner, P.E. (1989) “Expert decision making in evolving situations. *IEEE Transactions on Systems, Man, and Cybernetics, 19(3)*, pp. 606-615, 1989.

Appendix A: Questionnaire

Questions About Using the Templates (1=Strongly Agree, …, 5=Strongly Disagree)	**Average**
Together, the positive & negative evidence templates provide the basis for assessing whether a single sensor return supports or contradicts a hypothesis	**2.4**
The templates helped me to think of alternative causes that I would not have thought of otherwise	**2.2**
Thinking about the likelihood of each alternative cause helped me to evaluate the stated hypotheses	**1.8**
Thinking about how informed I was (expertise & data) helped me to evaluate the stated hypotheses	**1.8**
Overall the positive and negative evidence templates helped me to organize my thinking	**2.4**
It was easy to learn how to use the templates	**3**
The amount of time I took using the templates to interpret the cases was acceptable	**2.2**
The templates would help me communicate my interpretation process to other analysts	**2.2**
The templates provide a useful summary of a single sensor analysis that, if needed, could provide input into a larger sensor- fusion process	**2**
I would recommend that organizations responsible for interpreting single sensor returns use the positive and negative evidence templates I used today	**2.4**
Average	**2.24**