A System for Structured Intelligence Analysis and Its Evaluation

Tecuci G., Boicu M., Marcu D., Kaiser L.,
Tausczik, Y., Holincheck N.

Learning Agent Center, George Mason University

November 2020

Point of contact:
Dr. Gheorghe Tecuci, Department of Computer Science and Learning Agents Center
Email: *tecuci@gmu.edu*, *Tel: 703 993 1722, Fax: 703 993 1710*

Contents

Executive Summary

Cogent Argumentation with Crowd Elicitation (Co-Arg) is an analytical tool developed in the IARPA program Crowdsourcing Evidence, Argumentation, Thinking and Evaluation (CREATE). Co-Arg assists an intelligence analyst (called lead analyst) solve typical intelligence analysis problems with help from his/her colleagues (called crowd analysts), self-organized in an ad-hoc team. Developed in the framework of the scientific method, Co-Arg guides the lead analyst through a systematic process of evidence-based reasoning. From the questions s/he asks, s/he generates alternative hypotheses as possible answers to the questions. Then s/he puts each hypothesis to work to guide the discovery and collection of evidence by successively decomposing the hypothesis into simpler and simpler hypotheses that more clearly show what evidence is needed to prove or disprove them. This leads both to the construction of a tree-like Wigmorean argumentation structure and to the discovery of new evidence that can now be used to assess the probability of the hypotheses. This is accomplished by combining the credibility and relevance of the discovered evidence, based on an intuitive and easy to use symbolic probability system.

Through its two main components, Argupedia and Cogent, Co-Arg enables a synergistic integration of the lead analyst's imagination and expertise with computer's knowledge and critical reasoning, and the wisdom of the crowd analysts. Argupedia is a crowd problem-solving system enabling the crowd analysts to brainstorm in asynchronous sessions in order to solve tasks assigned by the lead analyst. It does this through a collection of easy to use, self-explanatory crowd tools that do not require significant training. In particular, Argupedia guides the crowd analysts to identify the competing hypotheses that answer the intelligence question asked, to develop informal arguments for each hypothesis, to identify the relevant evidence for each argument, and to assess its credibility. Cogent is an intelligent cognitive assistant that supports the lead analyst to evolve the informal analyses into detailed formal analyses that lay out the underlying analytic framework for every hypothesis, including the connection between the evidence and various intermediate conclusions in the analysis, the evaluation of the credibility of evidence and its strength in supporting a conclusion, and the role of any assumptions in addressing missing information. Cogent also identifies warnings, biases, and errors in the analysis, guiding the lead analyst in addressing them. It automatically updates the computed probabilities based on new or revised evidence, and it can automatically generate a structured report describing the analysis.

Co-Arg was evaluated in two experiments, an internal experiment conducted by the GMU-led team, pre-registered with the Center for Open Science, and an experiment conducted by the IARPA's testing and evaluation team (T&E) consisting of Good Judgement Inc. and John Hopkins University's Applied Physics Laboratory.

The internal experiment used a within-subjects design over the course of Spring and Summer 2018, with 62 students enrolled in four Universities. Participants worked on two test problems with Google Docs, then were trained in evidence-based reasoning and Co-Arg, and completed two test problems with Co-Arg. In both conditions, participants worked collaboratively in teams of 3-6 for 2.5 days, using Google Docs or Co-Arg depending on the condition, and then worked individually for 4.5 days, finalized their solution, wrote up the solution as a report, and submitted the report, using Google Docs or Co-Arg, depending on the condition. Reports were scored for quality of reasoning and quality of communication. The results showed that Co-Arg and its associated training improved quality of reasoning scores on intelligence problems by 0.77 points out of 10, improvement corresponding to a medium effect size of 0.48. A more detailed examination showed that the gains were due to improvement in the evaluation of sources of evidence and in the argumentation structure. We also found that on average, even when using Co-Arg, participants had poor quality of reasoning scores. With Google Docs on average participants scored 2.89 out of 10 and with Co-Arg, on average, participants scored 3.66 out of 10. Participants may have scored low on quality of reasoning for several reasons. First, the testing problems were very challenging. Second, the evaluation rubric was very detailed and it was very difficult to receive points for all of the items listed in the rubric. Third, participants were undergraduate and graduate students, many of whom had no training in intelligence analysis and no interest in intelligence analysis. Thus, compared to the target users, intelligence analysts, these users were less experienced and less motivated. Fourth, using either system, participants consistently scored very low in terms of one of the dimensions: identification of key missing information and assumptions. The current version of Co-Arg does not assist users with this type of reasoning. It is clear from this study that tools and training are needed to help users with identifying missing information and assumptions.

The results also showed that Co-Arg and its associated training improved scores for quality of communication by 1.04 points out of 6, improvement corresponding to a large effect size of 1.11. On average, students scored 2.86 out of 6 on quality of communication when solving intelligence problems with Google Docs, and 3.90 out of 6 when solving intelligence problems with Co-Arg. When using Co-Arg, on average, reports had a main conclusion stated up front, were mostly coherent and organized, and had several clear ideas. In comparison, when using Google Docs, reports were on average less likely to have a main conclusion stated up front, were less likely to be well organized, did not make the reasons favoring and disfavoring the main conclusion clear enough, and did not explain the evidence as well.

Finally, we evaluated the usability of Co-Arg by using the System Usability Scale. Co-Arg scored 56.3 on average, slightly above the hypothesized score of 55. This demonstrates that participants in our internal experiment considered the system as reasonably easy to use.

The T&E experiment used 20 problems, two of them proposed by GMU, “Fillistan Conducts Ballistic Missile Tests” in Round 1, and “Who is the Spy?” in Round 2. In our view, these were the two most difficult problems of all the problems that were for use in the T&E experiment. Both of them meet all the 8 key elements of high-quality analytic reasoning identified by IARPA and T&E, including “generation of unique analytic insights.” The T&E experiment resulted in five solutions with Co-Arg of the “Fillistan Conducts Ballistic Missile Tests” problem and four solutions with Co-Arg of the “Who is the Spy?” problem. We rated one solution for each problem. The solution of the missile problem was one of the best that we have seen for this difficult problem, including in our internal testing. It scored 31.5 points out of 39 (8.08 on a 10 point scale). The solution of the spy problem scored 29 out of 47 points (6.17 on a 10 points scale). These solutions represent a proof of concept that Co-Arg can be used to solve complex problems with a limited amount of training. The dexterity that participants demonstrated in using Co-Arg after only two hours of required training was impressive.

These results also suggest that our assumption that students would be motivated by the class grade to diligently learn and use Co-Arg may be wrong. We provided five practice problems to these students, and all were required. However, on average, these participants fully completed only 2.4 practice problems. The participants in the T&E experiment appear to have been quite motivated, and obtained better results than the students who had much more training.

1. Introduction

There is a huge gap between the ability to collect information and the ability to analyze it. In an effort to reduce this gap, IARPA launched the Crowdsourcing Evidence, Argumentation, Thinking and Evaluation (CREATE) program, an effort to develop and experimentally test systems that use crowdsourcing and structured analytic techniques to improve analytic reasoning (IARPA, 2016).

This section reviews the fields of structured analytic techniques and crowdsourcing, and introduces the teams that participated in the CREATE program and their approaches. Then, Section 2 presents the analytic framework of the Co-Arg system that is described in Section 3. Section 4 presents the strategies for managing the users of Co-Arg and Section 5 presents the training requirements for these users. Section 6 introduces the problems developed to evaluate Co-Arg. Section 7 presents the internal experimentation and evaluation of Co-Arg. Finally, Section 8 presents the use of Co-Arg in the experimentation conducted by the IARPA testing and evaluation team.

1.1. Structured Analytic Techniques

The complexity of intelligence analysis has led to a significant amount of research on developing structured analytic techniques (SATs) and computer-based tools to assist analysts. The main tradeoff faced by these methods and tools is between being simple and easy to use, on one hand, and providing significant assistance with all the complexities of the analytic process, on the other hand.

The majority of current SATs are used as independent methods, unlinked to an existing formal analysis. The most favored method by analysts is Analysis of Competing Hypotheses (ACH) introduced by Heuer (1999; 2008) which facilitates comparison of alternative hypotheses with respect to their evidence. ACH is at the “simple and easy-to-use” end of the spectrum, but the price paid is an overly-simplified view of the analysis process that provides only limited assistance. Efforts have been made to extend the basic ACH method with a very simple argumentation (Wheaton and Chido, 2006), with Bayesian probabilities (Valtorta et al, 2005), with symbolic probabilities (Pope and Josang, 2005), and with collaborative analysis (Heuer and Pherson, 2011, pp.165-169), all with limited success in addressing the full complexity of the analytic process.

More complex methods and tools developed are graphical editors for structured argumentation, best known being Rationale (2016; van Gelder, 2007) which is based on Toulmin’s argumentation model (1963). Rationale clearly shows both the reasons that support a hypothesis and those that oppose it, and whether they are based on evidence or on assumptions. Although it allows the analyst to attach symbolic probabilities to evidence and hypotheses, it does not rely on a probabilistic system, nor does it assist the analyst to develop the argumentation.

But analyst’s conclusions, which rest on evidence, are necessarily probabilistic in nature. Thus, any tool which is not based on a rigorous probabilistic system, including ACH and Rationale, is of limited utility.

Among the most complex analytic tools developed are those based on Subjective Bayesian reasoning, such as Netica (2016) and Hugin (2016). The mathematical underpinnings of this probabilistic system led to a very rich system for capturing a very wide array of evidential and inferential subtleties or complexities (Schum, 2001, Chapters 6, 7, and 8). The Bayesian system incorporates the concept of conditional dependence that provides the primary means for capturing evidential and inferential complexities for study and analysis. One distinct virtue of the Bayesian analysis of evidence is that it prompts us to ask questions of our evidence that we might never have thought of asking. But the Bayesian methods also have some important deficiencies, especially for intelligence analyses where most of the events of concern are unique, singular, or one-of-a-kind, so there is nothing to count. This also means that different analysts may assess these probabilities differently and arrive at different probabilities regarding major conclusions.

One problem with any analytic tool that relies on a probabilistic inference network is the difficulty of building the network, and the Bayesian networks are notoriously hard to build. This is the main reason why these analytic tools are rarely used by intelligence analysts, despite their power.

For several years we have worked on a computational theory of intelligence analysis (Tecuci et al., 2011) and, on this basis, we have developed a sequence of increasingly more practical cognitive assistants for the intelligence analysis education and practice. The first of these systems, Disciple-LTA (Tecuci et al., 2005; 2007b; 2008), is a unique and complex cognitive assistant that integrates powerful capabilities for analytic assistance, learning and tutoring, and is at the basis of the other developed systems. TIACRITIS (Teaching Intelligence Analysts Critical Thinking Skills) was developed primarily for teaching intelligence analysis and was experimentally used in many IC and DOD organizations (Tecuci et al., 2011). While praising its solid theoretical framework and deep evidentiary knowledge, the analysts desired a simplified interface and interaction. The next system, Disciple-CD (Disciple cognitive assistant for Connecting the Dots) significantly improved TIACRITIS along several dimensions (e.g., use of the Baconian probability system, easier argument development, more flexible management of knowledge bases, improved usability and scalability), and is accompanied by a textbook on Intelligence Analysis (Tecuci et al, 2014; 2016b). In the latest system, Cogent, with significant feedback from intelligence analysts, the user experience was significantly improved while preserving the sound foundations in the computational theory of intelligence analysis (Tecuci et al., 2018).

1.2. Wisdom of the Crowds

Crowdsourcing is a process through which one may obtain goods or services from a group of users usually over the Internet. This group is assumed to be large, open, and dynamic. The work is divided between participants, often through simple, repetitive tasks and then cumulated in order to obtain the desired result. While forms of crowdsourcing took place before, the term was coined around 2005 (Safire, 2009; Howe, 2006). We will briefly present some successful forms of crowdsourcing.

The most popular crowdsourcing initiative, Wikipedia (Reagle, 2010) with its goal of creating a public encyclopedia, demonstrated the great potential of this approach but also its drawbacks (including vandalism, hoaxes, and scandals) and management difficulties, such as “crushing bureaucracy,” new editors’ retention, monolithic administration and policies, and male bias (Simonite, 2013).

The Stanford Encyclopedia of Philosophy (SEP, 2016) alleviates some of these difficulties through an innovative competence-based editorial process (called dynamic reference work) that relies on the wisdom of a crowd of domain experts to provide the “most authoritative, comprehensive, and up-to-date information” resource on specific philosophy topics (Sonnad, 2015).

TopCoder (Lakhani et al., 2010) is creating a community of designers and technologists for software development. A notable aspect is the use of a competition-based approach in some of the projects developed. Clients include NASA, IARPA, IBM, and Harvard Medical School.

When a crowd-oriented validated evaluation method is used, crowd workers may perform comparably with experts in the assessment of very specialized tasks. For example, crowd workers on Amazon's Mechanical Turk and Facebook, without surgery-related experience, were able to assess robotic surgical suturing performance at a level comparable to that of experienced surgeons (Chen et al., 2013). Many other crowd selection approaches for quality assurance have been proposed, including using a skill ontology-based model (Maarry, 2014), a hierarchical taxonomy of skills (Mavridis, 2016), and the analysis of their social activities, relationships, and shared content (Bozzon et al., 2013).

1.3. The CREATE Teams and their Approaches to Improve Reasoning

There were four performers in CREATE, each led by a university (George Mason University, Monash University, Syracuse University, and University of Melbourne), and each investigating a different approach.

George Mason University has developed Co-Arg, a cognitive assistant based on a theory of evidence-based reasoning with Wigmorean arguments (Schum, 1987; 2001; Tecuci et al., 2016a, 2016b). Co-Arg supports answering intelligence questions by synergistically integrating an analyst's imagination and expertise with computer's knowledge and critical reasoning, and the wisdom of the crowd.

Monash University has developed BARD (Bayesian Argumentation via Delphi), a system that answers intelligence questions by using causal Bayesian networks as underlying structured representations for argument analysis and automated Delphi methods to bring groups of analysts to a consensus analysis.

Syracuse University has developing TRACE (Trackable Reasoning and Analysis for Crowdsourcing and Evaluation), the goal of which was to experimentally evaluate existing structured analytic techniques in order to determine the most effective ones.

Finally, University of Melbourne has developed SWARM (Smartly-assembled Wiki-style Argument Marshalling), an online collaboration platform supporting evidence-based reasoning by cultivating user engagement, exploiting natural expertise, and supporting rich collaboration.

There were also two control systems based on Google Docs, Conclude (for an individual user) and Concur (for a team).

All these six systems were integrated into a common evaluation environment developed by the testing and evaluation team (T&E) consisting of Good Judgement Inc. and John Hopkins University's Applied Physics Laboratory.

2. Analysis in the Framework of the Scientific Method

2.1. Computational Theory of Intelligence Analysis

Co-Arg builds directly on GMU's research on developing a computational theory of intelligence analysis as a basis for building advanced cognitive assistants for intelligence analysis education and practice. This theory is described in (Tecuci, et al., 2016b). Developed in the framework of the scientific method, this theory views intelligence analysis as a continuous collaboration of three processes, *evidence (or questions) in search of hypotheses*, *hypotheses in search of evidence*, and *evidentiary assessment of hypotheses*, performed jointly by an intelligence analyst and his or her cognitive assistant that incorporates a significant amount of knowledge from the science of evidence (see Figure 1). From the observations they make or the questions they ask, they generate alternative hypotheses as explanations of the observations or answers to the questions. Then they put each hypothesis to work to guide the discovery of evidence by decomposing them into simpler hypotheses that more clearly show what evidence is needed to prove or disprove them. This leads both to the discovery of new evidence and to the construction of a tree-like Wigmorean argumentation structure that can now be used to assess the probability of the hypothesis, based on the discovered evidence. This is a recursive process where, for example, the discovery of new evidence leads to the generation of new hypotheses or the modification of the existing ones which, in turn, lead to the discovery of new evidence.

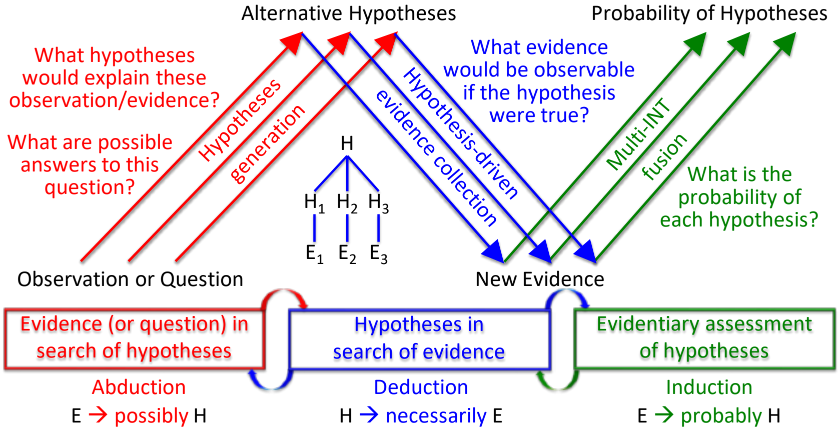

Alternative Hypotheses
Probability of Hypotheses
What hypotheses would explain these observation/evidence?
What are possible answers to this question?
Hypotheses generation
What evidence would be observable if the hypothesis were true?
Hypothesis-driven evidence collection
Multi-INT fusion
What is the probability of each hypothesis?
H
H1 H2 H3
E1 E2 E3
Observation or Question
New Evidence
Evidence (or question) in search of hypotheses
Hypotheses in search of evidence
Evidentiary assessment of hypotheses
Abduction
E → possibly H
Deduction
H → necessarily E
Induction
E → probably H

Figure 1: Analysis framework grounded in the scientific method.

Evidentiary assessment of hypotheses is probabilistic in nature because the evidence is always *incomplete* no matter how much we have and is commonly *inconclusive* in the sense that it is consistent with the truth of more than one hypothesis. Further, the evidence is frequently *ambiguous*, with multiple meanings. A mass of evidence is in most situations *dissonant*, some favoring and some disfavoring the hypothesis under consideration. Finally, the evidence comes from sources with different levels of *credibility*. One difficulty is that none of the most-studied non-enumerative probability views (i.e., Subjective Bayes, Belief Functions, Fuzzy, and Baconian) can optimally cope with all these evidence characteristics. For example, the *Subjective Bayesian* view cannot optimally cope with ambiguities or imprecision in evidence. On the other hand, the *Belief Functions* view (Shafer, 1976; Schum, 2001) and the *Fuzzy* view (Zadeh, 1983; Negoita and Ralescu, 1975) can naturally cope with imprecisions in evidence. The *Baconian* view, where the probability of a hypothesis depends on how many evidentiary tests the hypothesis has passed, is the only probability view that can optimally deal with the incompleteness of evidence (Cohen, 1977; 1989). The Subjective Bayesian, Belief Functions, and Fuzzy views all answer the question: *How strong is the evidence we do have about this hypothesis?* It is thus possible to encounter a situation where, based on the current evidence, all these three views predict that H_3 is the most likely hypothesis, just to learn later that H_1 was the true one. The Baconian view would have helped with this situation because it answers the question: *How much evidence do we have about this hypothesis, and how many questions about it remain unanswered?* Clearly, in the invoked situation, the answers to these unasked questions did not favor H_3. While on the Bayesian probability scale "0" means *disproof*, on the Baconian scale, "0" simply means *lack of proof*. A hypothesis now having "0" Baconian probability can be revised upward in probability as soon as we have some evidence for it. But we cannot revise upward in probability any hypothesis disproved, or having "0" conventional probability (Tecuci et al., 2016b).

A probabilistic system based on ideas from both the Baconian view and the Fuzzy view may potentially cope with all the five characteristics of evidence. Moreover, the use of similar min/max probability combination rules by the Baconian and the Fuzzy views facilitates the development of such an integrated system (Schum, 2001; Tecuci et al, 2016b; Cohen, 1977; 1989; Zadeh, 1983). These rules are much simpler than the Bayesian probability combination rules, which is important for the understandability of the analysis.

There is also the issue of using a numerical probability scale, which is required by a Bayesian view, as opposed to a symbolic scale required by a Fuzzy view. While a numerical probability is much more precise, it is not at all clear how an analyst would be able to defend a subjective assessment that, for instance, might assess the probability of a hypothesis H_k as exactly 77%. Analysts would arrive at different probability assessments, which would impede their collaboration. Because words are less precise than numbers, there will often be less disagreement about a verbal or fuzzy probability.

Starting from such considerations, we have defined an intuitive and easy to use system of *Baconian probabilities* with *Fuzzy qualifiers*. This system uses the following positive probability scale that is a refinement of the scale provided in the Intelligence Community Directive 203 (2007):

> lacking support (0-50%) < barely likely (50-55%) < likely (55-70%) < more than likely (70-80%) <
> very likely (80-95%) < almost certain (95-99%) < certain (100%).

If the evidence does not support the truthfulness of the hypothesis **H** (i.e., **H** is lacking support), then it may support the truthfulness of its negation, **not H**. In such a case, the probability of **H** may be expressed using the following negative probability scale: no chance (0%) < almost no chance (1-5%) < very unlikely (5-20%) < more than unlikely (20-30%) < unlikely (30-45%) < barely unlikely (45-50%).

2.2. Evidence-based Reasoning with Cogent and Co-Arg

In the following we provide the basic elements of the systematic approach to evidence-based reasoning to be followed when using the Cogent component of Co-Arg. This is part of what the users of Co-Arg are being taught before they can properly use this system. We include it here in order to better understand this report.

The first step in answering an intelligence question is to imagine potential answers. Each such answer is a hypothesis to be assessed. You need to determine which hypothesis is best supported by the available information.

To assess the probability that a hypothesis is true you have to develop an argumentation. *An argumentation is a reasoning structure that shows how the evidence and our assumptions support or refute the hypothesis.*

Consider, for example, the hypothesis "Hakka has chemical weapons." One way to support it is to show that Hakka, which is an apocalyptic sect, develops chemical weapons. This, in turn, would be supported by Hakka having the necessary expertise, production materials, and funds (see top part of Figure 2). Each of these sub-hypotheses need to be supported by evidence or by making assumptions.

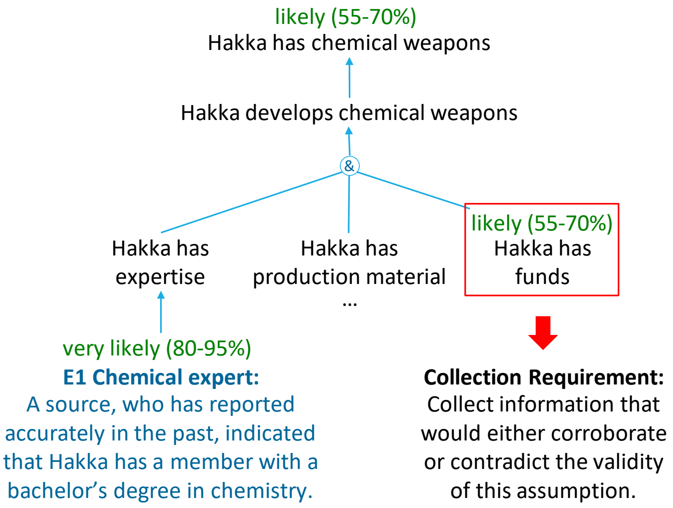

likely (55-70%)
Hakka has chemical weapons
Hakka develops chemical weapons
&
Hakka has expertise
Hakka has production material
…
likely (55-70%)
Hakka has funds
very likely (80-95%)
E1 Chemical expert:
A source, who has reported accurately in the past, indicated that Hakka has a member with a bachelor's degree in chemistry.
Collection Requirement:
Collect information that would either corroborate or contradict the validity of this assumption.

Figure 2: Simple Wigmorean argumentation.

Consider the following information: "A source, who has reported accurately in the past, indicated that Hakka has a member with a bachelor's degree in chemistry." This information supports the truthfulness of the hypothesis that "Hakka has expertise to develop chemical weapons," and is therefore evidence for this hypothesis. Let's name it: "E1 Chemical expert" (see Figure 2).

Evidence is any item of information that favors or disfavors the truthfulness of a hypothesis (Schum, 2001). We say that the evidence is relevant to that hypothesis.

It is also possible that, for some of the sub-hypotheses, such as "Hakka has funds," we may not have any evidence. In such a case we may treat it as an assumption. *An assumption is a statement taken to be true, based on knowledge about similar situations and commonsense reasoning, without having any supporting evidence.* For example, we may assume that it is *likely (55-70%)* that "Hakka has funds."

As a result of all these uncertainties, we will not be able to prove that the top hypothesis is definitely true or definitely false, but we will be able to estimate the probability of it being true or false, such as "It is *likely (55-70%)* that Hakka has chemical weapons."

One way of improving the quality of the analysis is to collect additional information that would either corroborate or contradict the validity of the assumptions made.

In the following, we are going to show how to answer questions based on imperfect information, by generating competing hypotheses and building argumentations for assessing which hypothesis is the most likely. We will start by introducing the main characteristics or credentials of evidence: *credibility*, *relevance*, and *inferential force*.

It is important to distinguish between *evidence about a fact* and the *fact* itself. Consider the item of evidence from Figure 2. Can we conclude from it that Hakka has a member with a bachelor's degree in chemistry? No. At issue here is the credibility of the source who may or may not be telling the truth. ***Credibility** of an item of evidence is the extent to which the evidence may be believed.* This assessment can be influenced by many things, including doubts about the source's veracity or by more credible information that contradicts this item of evidence.

We can assess the credibility of this item of evidence by answering the following question: *What is the probability that the evidence is true?* The source of this evidence has reported accurately in the past. We can therefore assume that this current report is very likely to be true.

Another credential or property of evidence is its relevance. *The relevance of an item of evidence indicates how strongly this item supports a specific hypothesis in the argument.* Relevance depends on how *recent* the evidence is, how *unambiguous* it is, and how *conclusive* the link between the evidence and the hypothesis is. The evidence may be unambiguous but it may support more than one hypothesis. We can assess the relevance by answering the question: *Assuming that the evidence is true, what is the probability that the hypothesis is true?*

When we have evidence about a fact and the hypothesis is the fact itself, the relevance of evidence is certain. Indeed, if we assume that the evidence is true, then the hypothesis is true (see the bottom part of Figure 3).

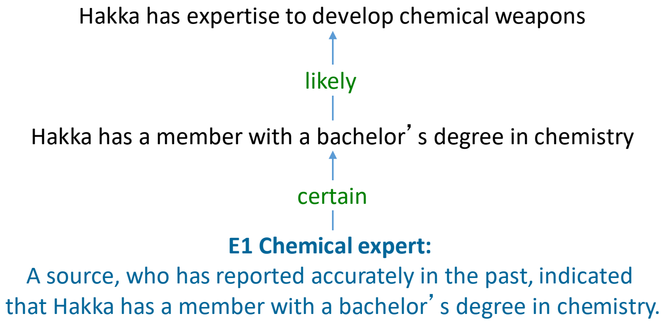

Hakka has expertise to develop chemical weapons
likely
Hakka has a member with a bachelor’ s degree in chemistry
certain
E1 Chemical expert:
A source, who has reported accurately in the past, indicated
that Hakka has a member with a bachelor’ s degree in chemistry.

Figure 3: Relevance of evidence and argument.

But let us consider the hypothesis that “Hakka has expertise to develop chemical weapons.” If Hakka has a member with a bachelor’s degree in chemistry, what is the probability that it has expertise to develop chemical weapons? A bachelor program in chemistry does provide the basic knowledge for chemical weapons development, but this does not necessarily prove that Hakka has the expertise. Indeed, the Hakka member may not have developed this expertise. Thus this item of evidence is not conclusive and we assess its relevance only as *likely* (see top part of Figure 3). Such explanations for assessments of relevance that are less than certain will clarify the reasoning and can be used in the completed argumentation to support the conclusion.

When developing an argumentation, it is a good practice to consider, for each item of evidence, which is the corresponding fact, and then reason from that fact to the upper-level hypotheses, as illustrated in Figure 3.

The third credential of evidence is its inferential force. Consider the inference from the bottom part of Figure 4. We have assessed the relevance of the item of evidence E1 as *certain* because it is an inference from evidence about a fact to the fact itself. We have also assessed the credibility of E1 as *very likely* because the source has reported accurately in the past. *Inferential force* answers the question: *What is the probability of the hypothesis above based only on this item of evidence below?* In our example, the relevance of E1 is *certain*, but its credibility is only *very likely*. Therefore the probability that “Hakka has a member with a bachelor’s degree in chemistry” is only *very likely*.

In general, the inferential force of an item of evidence is determined as the smaller between its credibility and its relevance. Indeed, an item of evidence that is not credible would not convince us that the hypothesis is true, no matter how relevant the provided information is. Therefore the inferential force in this circumstance would be low. Similarly, it is not enough for the item of evidence to be credible, if the information provided is not relevant to the hypothesis. The inferential force will be high only if the evidence item is both highly relevant and credible.

In this case, because we have only one item of evidence, the probability of the hypothesis "Hakka has a member with a bachelor's degree in chemistry" is given by the inferential force of this item of evidence. However, if we have more items of evidence, some favoring the truthfulness of the hypothesis, and some disfavoring it, then the probability of the hypothesis will result from the combined inferential force of all these items of evidence.

As another example, let's now consider the upper-level hypothesis from Figure 4: "Hakka has expertise to develop chemical weapons." The relevance of "Hakka has a member with a bachelor's degree in chemistry" to this hypothesis was assessed as *likely*. Because the probability of the sub-hypothesis is *very likely*, its inferential force to the top hypothesis is *likely*, the minimum of its relevance and probability.

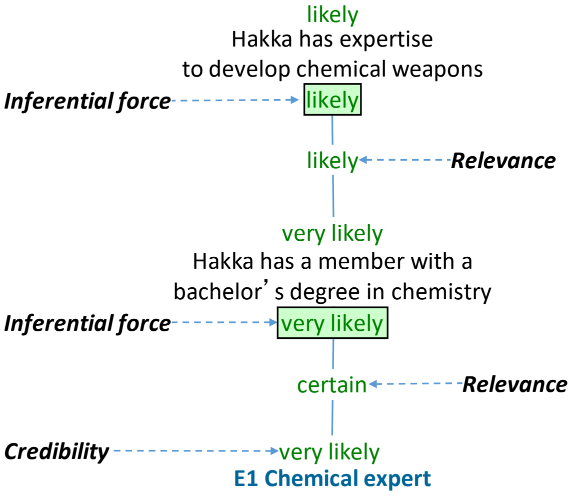


Figure 4: Credentials of evidence and arguments.

In this case we have just this one reason and one argument for the top hypothesis to be true. Therefore the probability of the top hypothesis is the same as the inferential force of this reason. In general, however, we may have multiple arguments, some favoring the truthfulness of the top hypothesis and some disfavoring it. In such a case the probability of the top hypothesis will be given by the combined inferential force of all these arguments.

3. Co-Arg: Cogent Argumentation with Crowd Elicitation

3.1. Co-Arg Architecture and Workflow

The overall architecture of Co-Arg, as integrated into the T&E evaluation environment, is shown in Figure 5. Co-Arg receives the schedule, users, and problems from the T&E Create Better Reasoning Portal and returns the reports developed by the users back to the T&E portal. A similar architecture was used in the internal evaluation, with the difference that the schedules, users, and problems were managed in Argupedia. The final reports were also submitted to Argupedia.

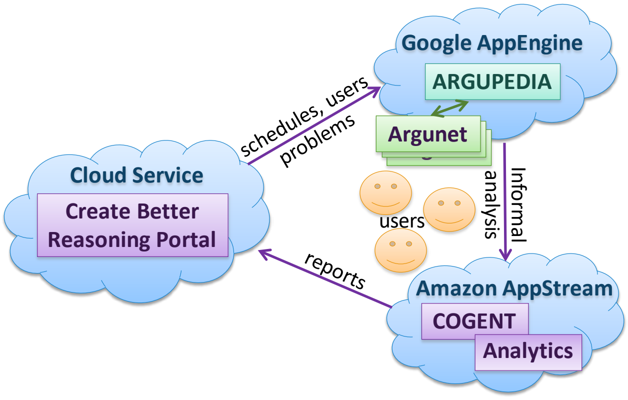
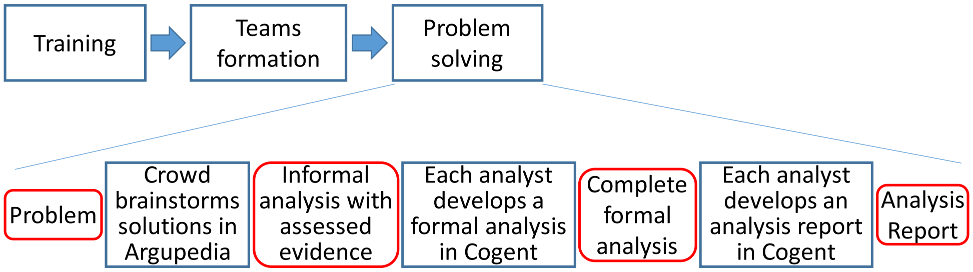


Figure 5: Co-Arg in the T&E evaluation environment.

We have designed and implemented the Co-Arg workflow from Figure 6 that enables the crowd to contribute to the analysis. First the users are rapidly trained in evidence-based reasoning and the use of Co-Arg. Then they are placed in teams, and each team follows the process from the bottom of Figure 6.



Training
Teams formation
Problem solving
Problem
Crowd brainstorms solutions in Argupedia
Informal analysis with assessed evidence
Each analyst develops a formal analysis in Cogent
Complete formal analysis
Each analyst develops an analysis report in Cogent
Analysis Report

Figure 6: Co-Arg workflow.

Argupedia receives a problem to be solved and the entire team uses it to asynchronously brainstorm possible answers to the intelligence question(s) asked. Then, for each imagined possible answer or hypothesis, the team collaborates in developing a brief informal argumentation expressed as a natural language phrase. Finally, the crowd attaches favoring and disfavoring evidence to the defined informal arguments and assesses the credibility of evidence.

The informal argumentations are passed to Cogent where each user, independently of the other users in the team, develops a formal analysis that completely lays out the argumentation for every hypothesis. Cogent detects biases and potential errors, guiding the user in addressing them. It facilitates the analysis of what-if scenarios, and automatically updates the analysis based on new or revised evidence. It can guide the user to perform deep credibility analysis of evidence. Finally, it generates a structured report that answers the intelligence question(s) asked.

3.2. Development of Informal Analysis with Argupedia

At the beginning of the analysis process each team member is instructed to independently read the description of the problem and to imagine potential hypotheses, arguments to support or refute them, and what information supports these arguments. We introduced this step following various internal tests in which some team members complained about the framing bias (i.e., a member of the team is influenced by how the members that have already started the brainstorming framed the problem).

Once familiarized with the problem, and having a personal opinion about the solution, the members of the team may start the asynchronous collaborative brainstorming process.

3.2.1. Starting the Informal Analysis

We illustrate the asynchronous collaboration process with a case study inspired by the actual use of the system in our internal experiments to solve the "Economics Minister Resigns" problem, shown in Table 1.

Situation:

The country Packland has two major political parties, the Home First Party (HFP) and the United Party (UP). Mark Gaines, the President of Packland, is up for reelection in October 2017 as the HFP candidate.

In 2014, the year before Jeremy Handle became Economics Minister, Packland's economic growth was 6 percent and unemployment was 3 percent; inflation, however, was running at 15 percent. He immediately implemented policies to lower inflation and within 12 months the inflation rate had fallen to about 6 percent, and it has remained in this range ever since.

Packland's economic growth slowed as a result of Handle's tight monetary policies. During 2015-16, the economy grew at an annual rate of 4 percent, according to statistics from the Economics Ministry, but during the first quarter of 2017, economic growth slowed to 1.8 percent. On 15 July 2017, Economics Minister Handle abruptly resigned. No explanation was given.

Question: Why did Handle resign or was asked to resign?

Available Information:

After several scandals involving government officials, in December 2016 Gaines publicly implemented an "honor code" in which members of his government were expected to behave in a upright and proper manner at all times. According to the new government policy on behavior, failure to adhere to this "honor code" would result in the dismissal of the government official irrespective of the official's position.

During several public speeches last year and in early 2017, Handle said that there was no need to pursue policies that risked higher inflation as long as inflation remained above 5 percent, according to newspaper accounts of these speeches.

Handle suffered a major heart attack in mid-January 2017. The president's office publicly announced that Handle would be taking a four-week leave of absence.

In early March 2017, Packland's energy minister, while intoxicated, was involved in a car accident that killed two brothers. The energy minister remained in the government but was forced to seek alcohol-abuse counseling. All of this information was available to the public and documented by court records.

In mid-February, the Economics Ministry publicly announced that Handle had returned to work.

In April 2017, a reliable source close to Handle said that Handle would oppose any change in monetary policies as long as economic growth was positive.

According to a poll of likely voters in May 2017, 30 percent of the respondents said they would vote for Gaines while 45 percent indicated their preference for the UP candidate; 25 percent were undecided. Those favoring the UP candidate listed dissatisfaction with the slowing economy as their primary reason for doing so. In a poll in July 2016, 53 percent of the respondents indicated they would likely vote for Gaines. A highly respected polling firm conducted the poll.

In early April 2017, a vice deputy defense minister John Habit, responsible for naval-related procurement, was convicted of soliciting a prostitute, according to court records.

In May 2017, a department head in the Defense Ministry responsible for procuring replacement parts for Air Force reconnaissance planes was publicly fired—in accordance with the new "honor-code"—after he was arrested for spousal abuse.

According to a reliable source on Gaines' reelection committee, this was the third official who has been fired for enforcement of the "honor-code" since its announcement in February 2017. A junior diplomat in the Foreign Ministry and a new employee in the Transportation Ministry were also dismissed for unbecoming behavior.

In March 2017, a newspaper reported that Handle was employing undocumented immigrants as house servants. Handle claimed he was unaware that these individual were in Packland illegally. The newspaper also reported that Handle was having an affair with one of the undocumented immigrants. Handle apologized publicly and said this embarrassing behavior would not happen again.

During a meeting with Gaines on 23 May 2017, Handle argued in favor of tight monetary policies designed to limit inflation, according to a reliable source on Handle's staff.

In late May 2017, Handle went skiing with his wife, according to several witnesses at a ski resort.

In late May 2017, the Defense Ministry announced that John Habit was now the deputy defense minister responsible for all military procurement.

In a press conference on 4 June, the president's spokesperson, after being asked whether the government was going to relax its stringent monetary policies, was noncommittal.

In one 15-second news clip in early June, Handle appeared pale and weak.

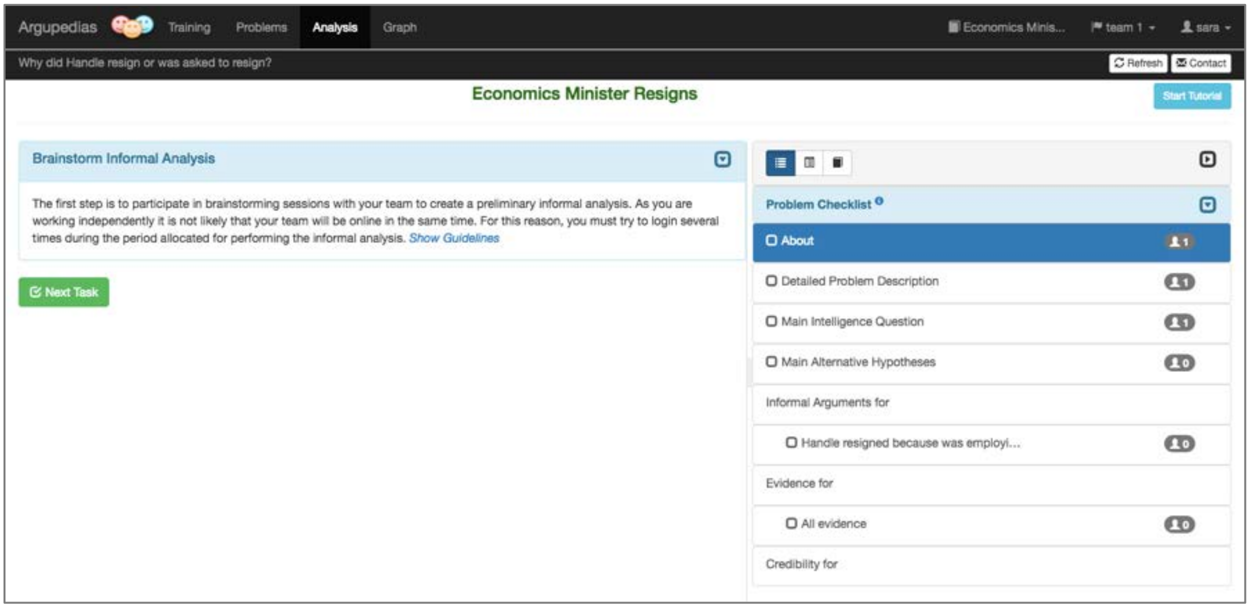
The reliable source on Handle's staff reported in mid-June that Handle was again working 60 hours a week.

In early July 2017, Gaines told the head of the HFP that he was now more concerned about the economy's slow growth than inflation. The source is an assistant to the HFP head.

In a press conference on 12 July, Gaines noted that he was disappointed in the economic results.

Table 1: The "Economics Minister Resigns" problem.

Let us imagine that Sara is a first-time user of the system. She just completed the training and then selected the "Economic Minister Resigns" problem to analyze. The home page of the informal analysis interface is shown in Figure 7.



Why did Handle resign or was asked to resign?

Economics Minister Resigns

Brainstorm Informal Analysis

The first step is to participate in brainstorming sessions with your team to create a preliminary informal analysis. As you are working independently it is not likely that your team will be online in the same time. For this reason, you must try to login several times during the period allocated for performing the informal analysis. *Show Guidelines*

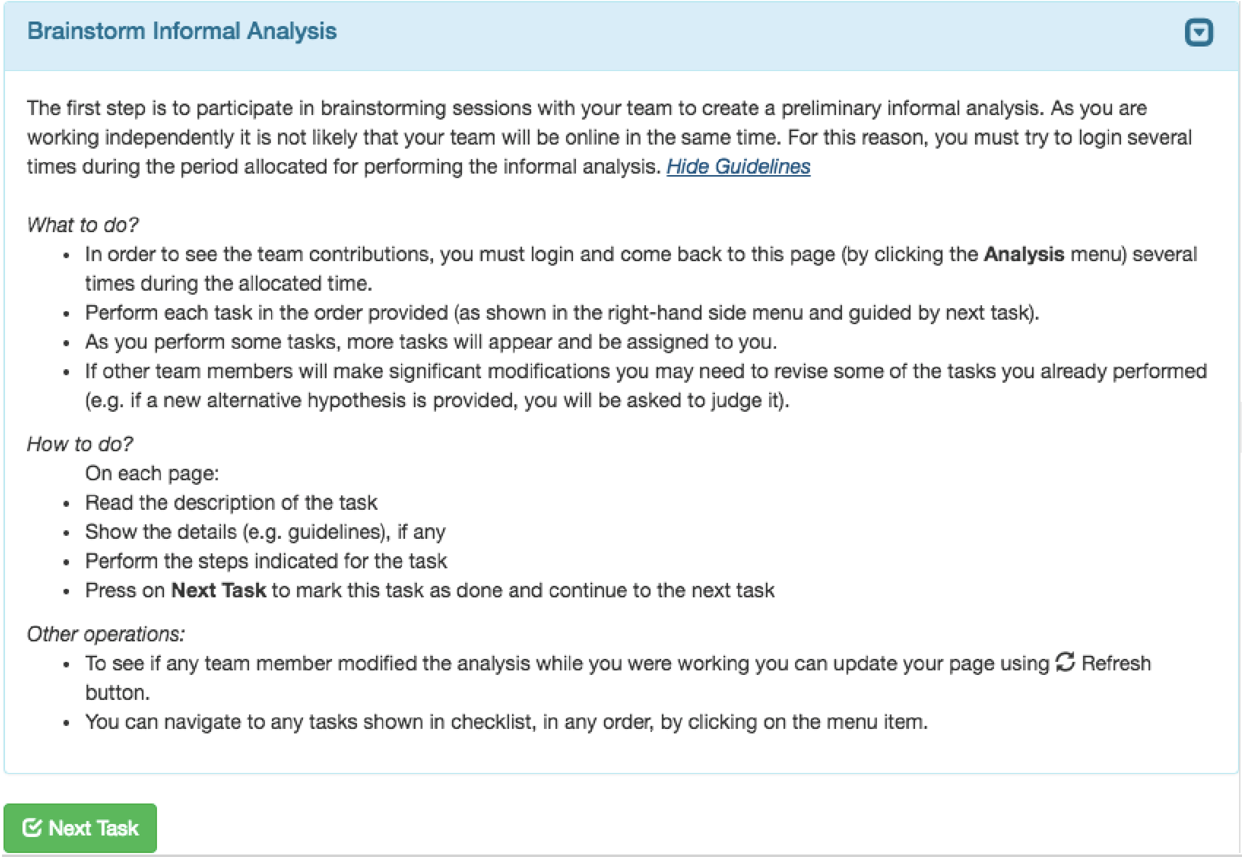
Problem Checklist

□ About	1
□ Detailed Problem Description	1
□ Main Intelligence Question	1
□ Main Alternative Hypotheses	0
Informal Arguments for	
□ Handle resigned because was employi...	0
Evidence for	
□ All evidence	0
Credibility for	

Figure 7: Informal analysis home page.

In the left-hand side of the interface, Sara will work with various tools, guiding her informal analysis. In the right-hand side there is a hierarchical description of the status of the informal analysis, which shows both her progress and the progress of her teammates.

All the steps in the informal analysis workflow start with a guide on what needs to be done. This guide shows, by default, a summary of the operation as a reminder. The user may request more details about some of the steps by clicking the Show Guidelines links. Sara, as a first-time user, desires to see how to operate the Analysis tool and expands the help provided, as shown in Figure 8. The detailed help contains more explanations related to what tasks the user needs to do, how to operate the tool, and what other operations are available to the user.



Brainstorm Informal Analysis

The first step is to participate in brainstorming sessions with your team to create a preliminary informal analysis. As you are working independently it is not likely that your team will be online in the same time. For this reason, you must try to login several times during the period allocated for performing the informal analysis. Hide Guidelines

What to do?

- In order to see the team contributions, you must login and come back to this page (by clicking the **Analysis** menu) several times during the allocated time.
- Perform each task in the order provided (as shown in the right-hand side menu and guided by next task).
- As you perform some tasks, more tasks will appear and be assigned to you.
- If other team members will make significant modifications you may need to revise some of the tasks you already performed (e.g. if a new alternative hypothesis is provided, you will be asked to judge it).

How to do?

On each page:

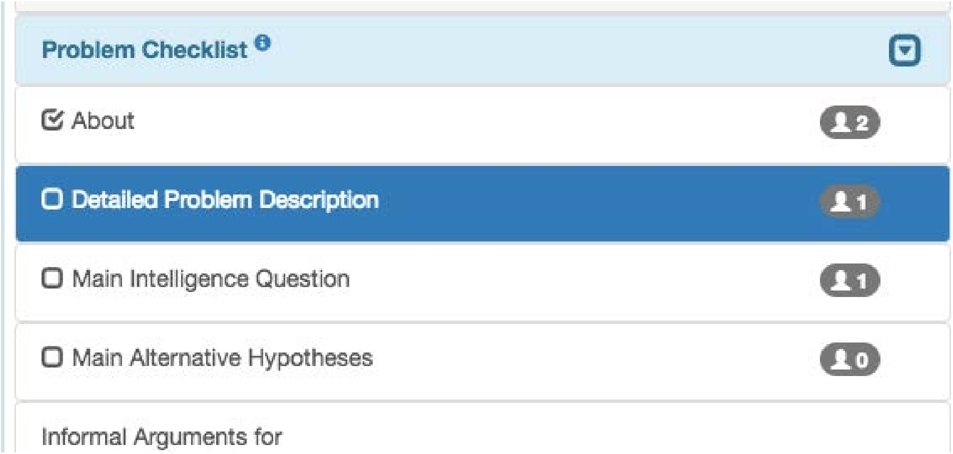
- Read the description of the task
- Show the details (e.g. guidelines), if any
- Perform the steps indicated for the task
- Press on **Next Task** to mark this task as done and continue to the next task

Other operations:

- To see if any team member modified the analysis while you were working you can update your page using Refresh button.
- You can navigate to any tasks shown in checklist, in any order, by clicking on the menu item.

Figure 8: General help for the informal analysis tool.

After reading the instructions, Sara clicks on **Next Task** button. The Problem Checklist component marks the progress, as shown in Figure 9.



Problem Checklist

☑ About 2

□ Detailed Problem Description 1

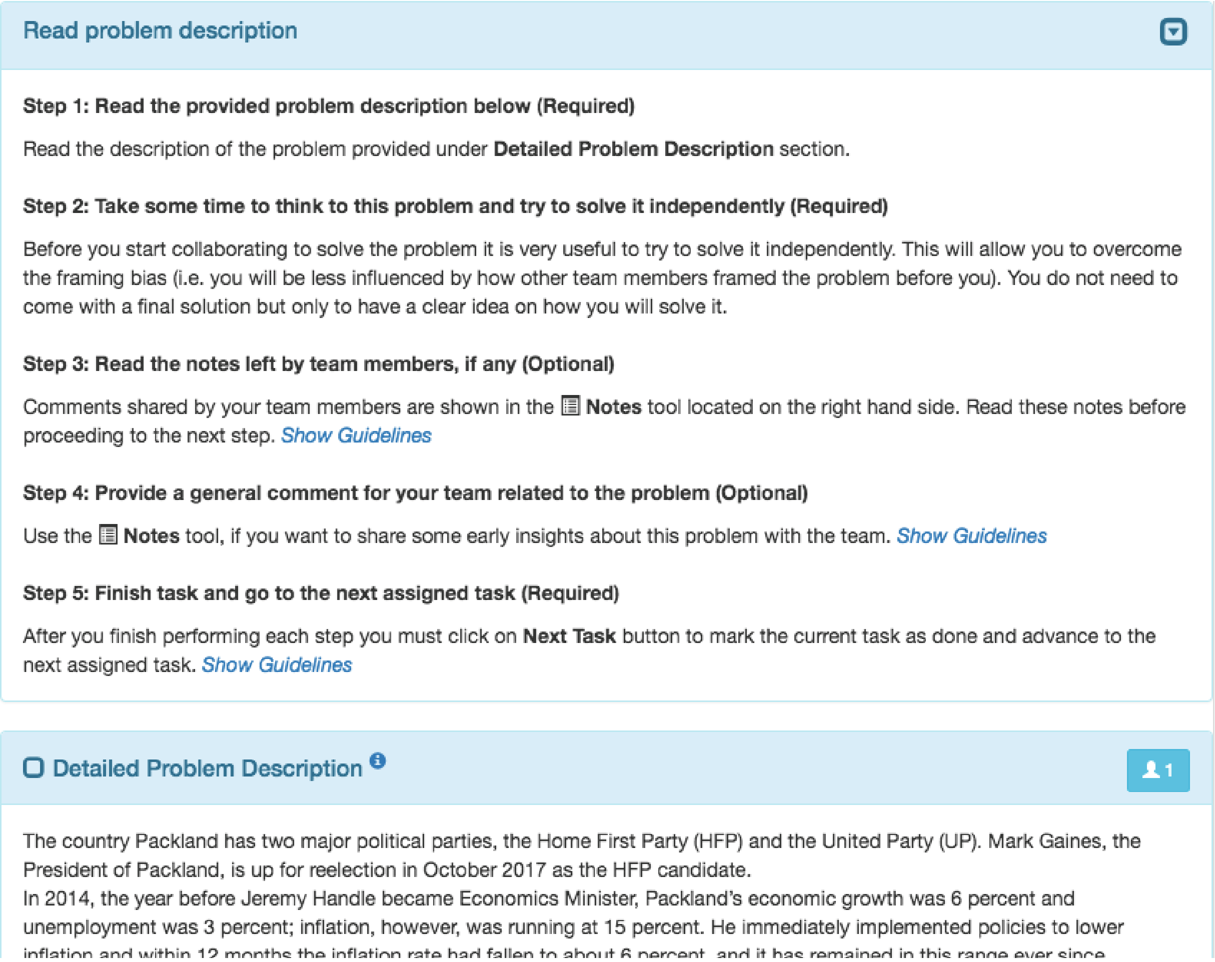
□ Main Intelligence Question 1

□ Main Alternative Hypotheses 0

Informal Arguments for

Figure 9: Tracking analysis progress in the Problem Checklist component.

The next task for Sara is to read the detailed problem description and, optionally, to provide general comments on the problem to the team.



Read problem description

Step 1: Read the provided problem description below (Required)

Read the description of the problem provided under **Detailed Problem Description** section.

Step 2: Take some time to think to this problem and try to solve it independently (Required)

Before you start collaborating to solve the problem it is very useful to try to solve it independently. This will allow you to overcome the framing bias (i.e. you will be less influenced by how other team members framed the problem before you). You do not need to come with a final solution but only to have a clear idea on how you will solve it.

Step 3: Read the notes left by team members, if any (Optional)

Comments shared by your team members are shown in the **Notes** tool located on the right hand side. Read these notes before proceeding to the next step. *Show Guidelines*

Step 4: Provide a general comment for your team related to the problem (Optional)

Use the **Notes** tool, if you want to share some early insights about this problem with the team. *Show Guidelines*

Step 5: Finish task and go to the next assigned task (Required)

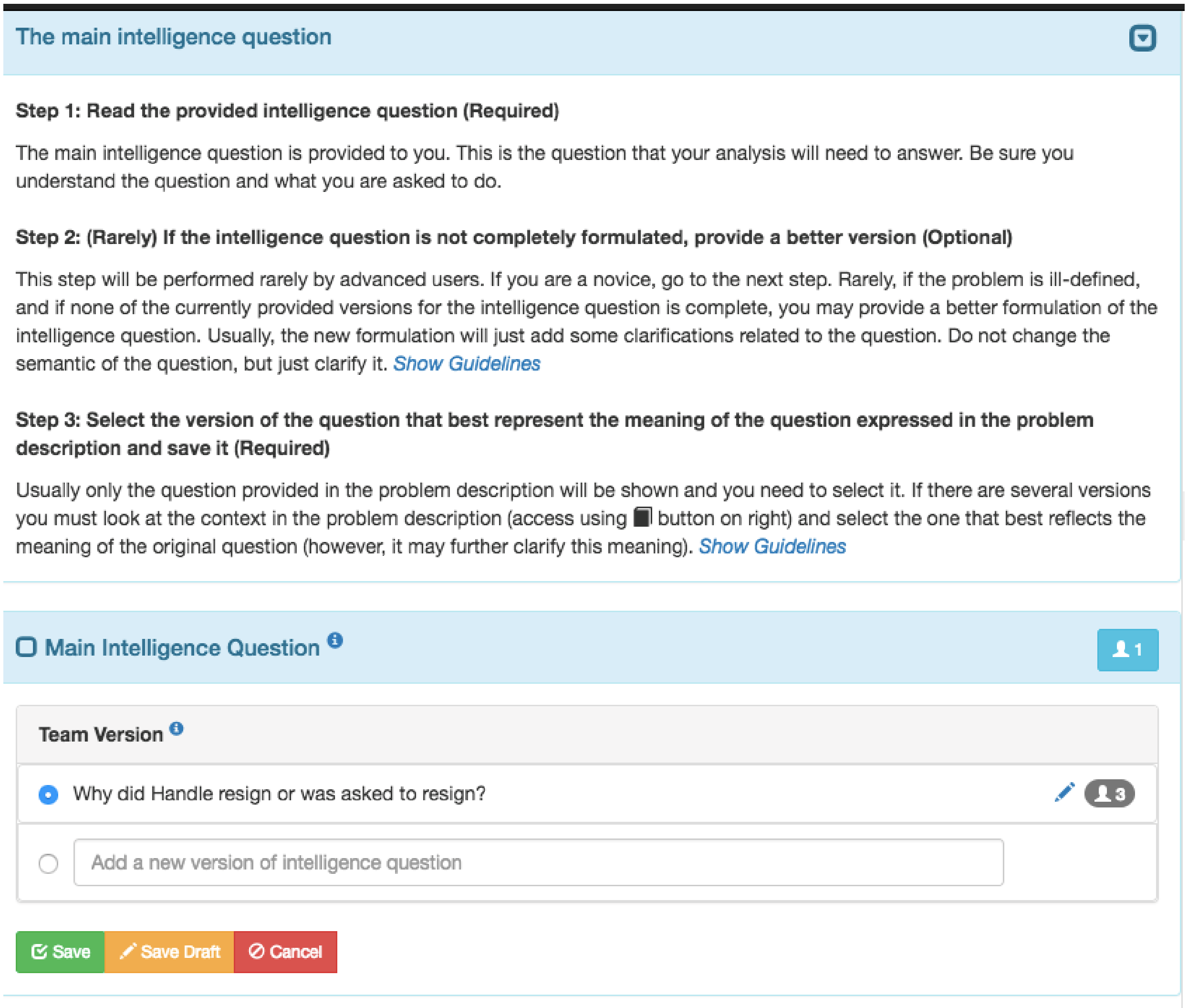
After you finish performing each step you must click on **Next Task** button to mark the current task as done and advance to the next assigned task. *Show Guidelines*

□ Detailed Problem Description 1

The country Packland has two major political parties, the Home First Party (HFP) and the United Party (UP). Mark Gaines, the President of Packland, is up for reelection in October 2017 as the HFP candidate.
In 2014, the year before Jeremy Handle became Economics Minister, Packland's economic growth was 6 percent and unemployment was 3 percent; inflation, however, was running at 15 percent. He immediately implemented policies to lower inflation and within 12 months the inflation rate had fallen to about 6 percent, and it has remained in this range ever since.

Figure 10: Detailed problem description.

After reading the problem description, Sara advances to the next task, the review of the intelligence question, as shown in Figure 11.



The main intelligence question

Step 1: Read the provided intelligence question (Required)

The main intelligence question is provided to you. This is the question that your analysis will need to answer. Be sure you understand the question and what you are asked to do.

Step 2: (Rarely) If the intelligence question is not completely formulated, provide a better version (Optional)

This step will be performed rarely by advanced users. If you are a novice, go to the next step. Rarely, if the problem is ill-defined, and if none of the currently provided versions for the intelligence question is complete, you may provide a better formulation of the intelligence question. Usually, the new formulation will just add some clarifications related to the question. Do not change the semantic of the question, but just clarify it. *Show Guidelines*

Step 3: Select the version of the question that best represent the meaning of the question expressed in the problem description and save it (Required)

Usually only the question provided in the problem description will be shown and you need to select it. If there are several versions you must look at the context in the problem description (access using ■ button on right) and select the one that best reflects the meaning of the original question (however, it may further clarify this meaning). *Show Guidelines*

□ Main Intelligence Question

Team Version

● Why did Handle resign or was asked to resign?

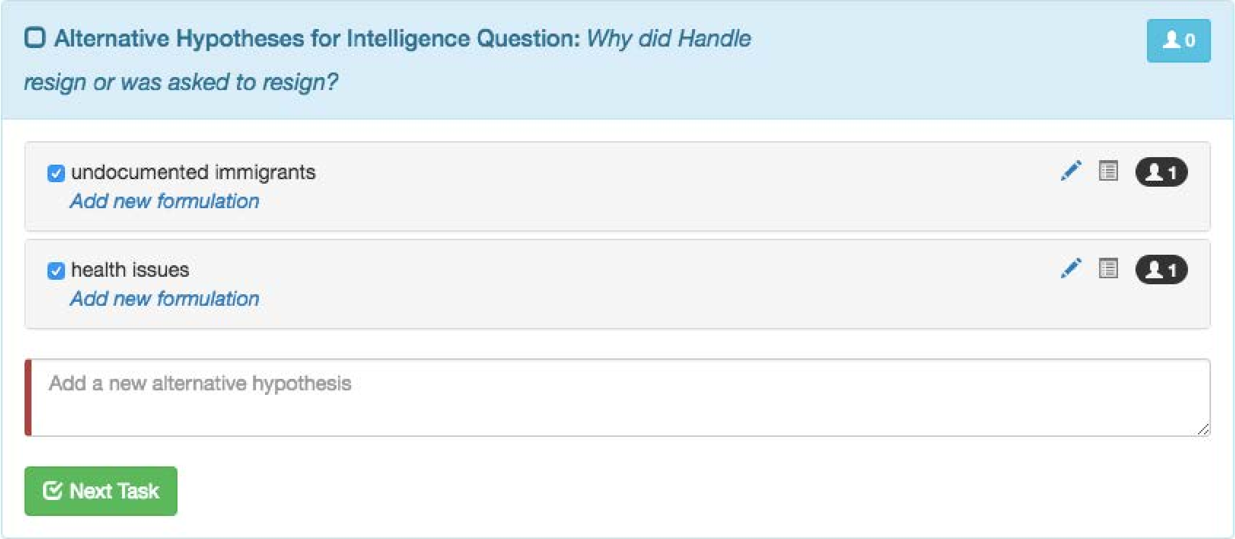
○ Add a new version of intelligence question

Figure 11: Intelligence question review.

The intelligence question asked is: "Why did Handle resign or was asked to resign?" For advanced users, this task allows a rephrasing of the intelligence question. Sara decides to keep the current form of the intelligence question, by clicking Save. Next, she wants to reflect on the problem, as recommended in the training materials, before continuing with the informal analysis.

3.2.2. Brainstorming Alternative Hypotheses

John is the next to enter the brainstorming process for informal analysis and, after performing the initial steps, provides two brief hypotheses as answers to the intelligence question: (1) undocumented immigrants and (2) health issues. Figure 12 shows the current crowd view for John’s entry. Notice in the right-hand side of the figure that only one person supports these hypotheses (John).



□ Alternative Hypotheses for Intelligence Question: *Why did Handle resign or was asked to resign?*

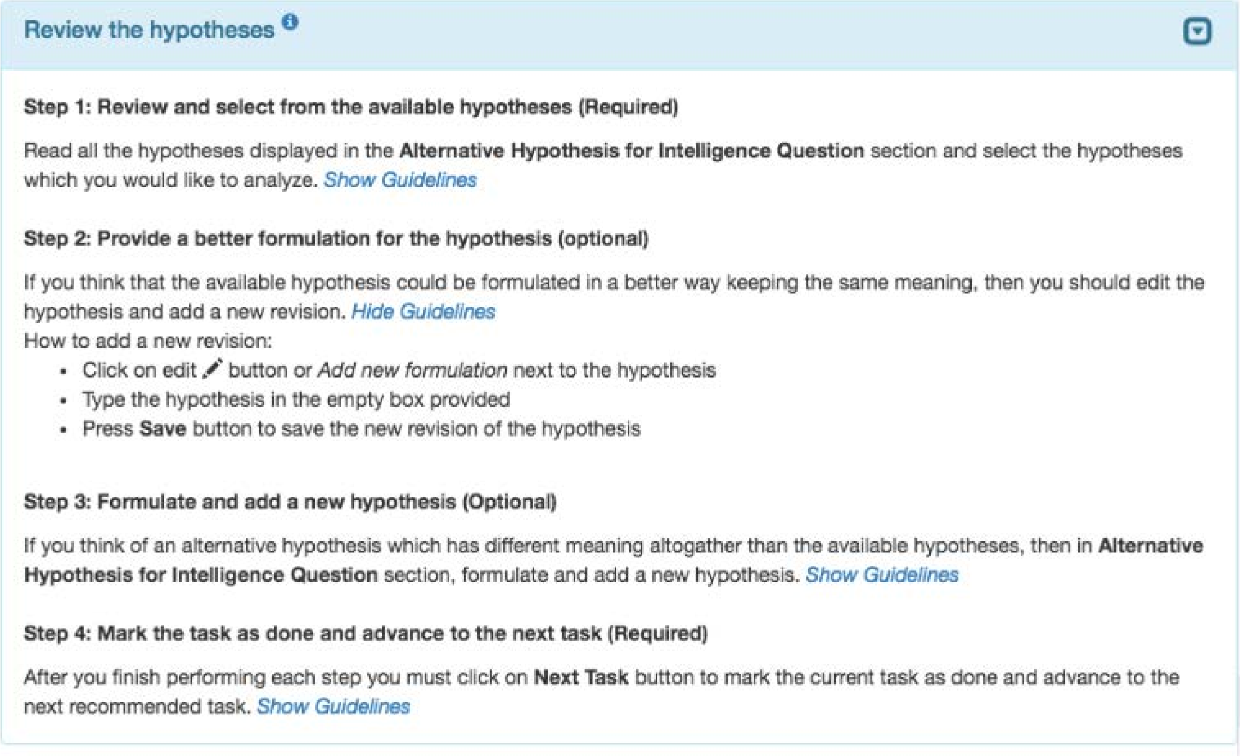
☑ undocumented immigrants
Add new formulation

☑ health issues
Add new formulation

Add a new alternative hypothesis

Figure 12: The initial contribution of analyst user in a brainstorming session.

After a while Mark joins the brainstorming and sees the hypotheses formulated by John. Although he agrees with them, he wants to improve their formulation, as instructed in the step by step guide (see Step 2 in Figure 13).



Review the hypotheses

Step 1: Review and select from the available hypotheses (Required)

Read all the hypotheses displayed in the **Alternative Hypothesis for Intelligence Question** section and select the hypotheses which you would like to analyze. *Show Guidelines*

Step 2: Provide a better formulation for the hypothesis (optional)

If you think that the available hypothesis could be formulated in a better way keeping the same meaning, then you should edit the hypothesis and add a new revision. *Hide Guidelines*
How to add a new revision:

- Click on edit ✎ button or *Add new formulation* next to the hypothesis
- Type the hypothesis in the empty box provided
- Press **Save** button to save the new revision of the hypothesis

Step 3: Formulate and add a new hypothesis (Optional)

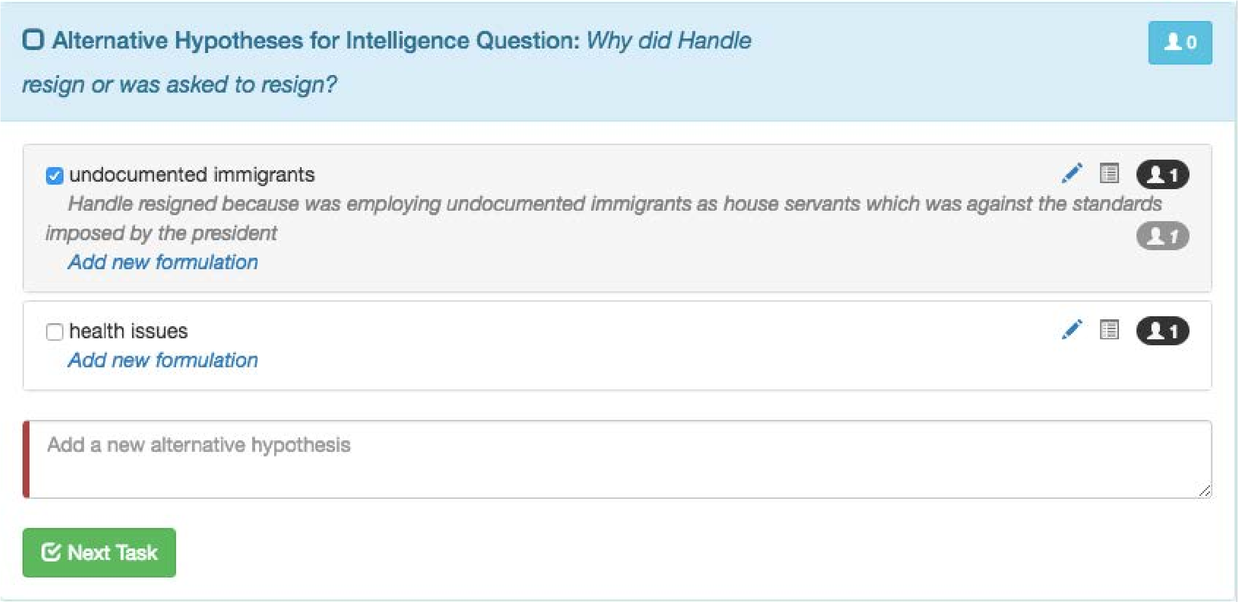
If you think of an alternative hypothesis which has different meaning altogather than the available hypotheses, then in **Alternative Hypothesis for Intelligence Question** section, formulate and add a new hypothesis. *Show Guidelines*

Step 4: Mark the task as done and advance to the next task (Required)

After you finish performing each step you must click on **Next Task** button to mark the current task as done and advance to the next recommended task. *Show Guidelines*

Figure 13: Step by step guidance for brainstorming alternative hypotheses.

Mark proposes a reformulation of the first hypothesis: Handle resigned because was employing undocumented immigrants as house servants which was against the morality standards imposed by the president. Figure 14 shows that now there are two versions for the first hypothesis, the initial version provided by John, and the revised version provided by Mark, each with a single vote (the vote icon has a tooltip that displays who voted for each version).



□ Alternative Hypotheses for Intelligence Question: *Why did Handle resign or was asked to resign?*

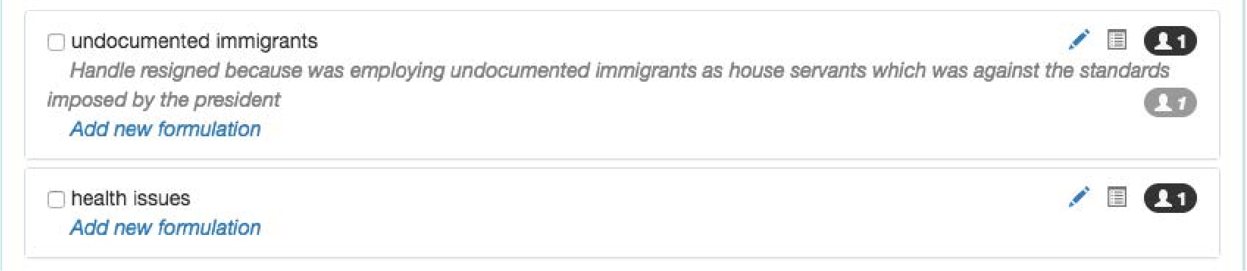
☑ undocumented immigrants

Handle resigned because was employing undocumented immigrants as house servants which was against the standards imposed by the president

□ health issues

Figure 14: Proposed reformulation for a hypothesis.

Now Mary joins the brainstorming and sees the hypotheses from Figure 15.



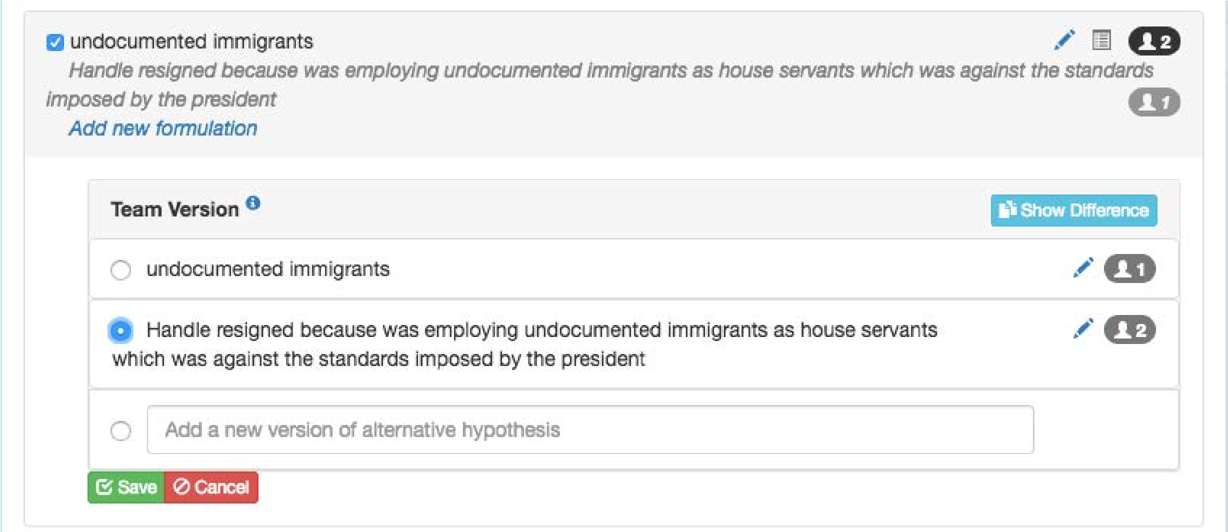
□ undocumented immigrants

Handle resigned because was employing undocumented immigrants as house servants which was against the standards imposed by the president

□ health issues

Figure 15: Brainstorming summary.

In the current implementation, at equal votes, the oldest version has priority. Therefore, John’s version is still shown as the team’s version. Mary decides to reformulate both hypotheses, starting with selecting the version provided by Mark for the first hypothesis (see Figure 16).


☑ undocumented immigrants 2

Handle resigned because was employing undocumented immigrants as house servants which was against the standards imposed by the president 1

Team Version

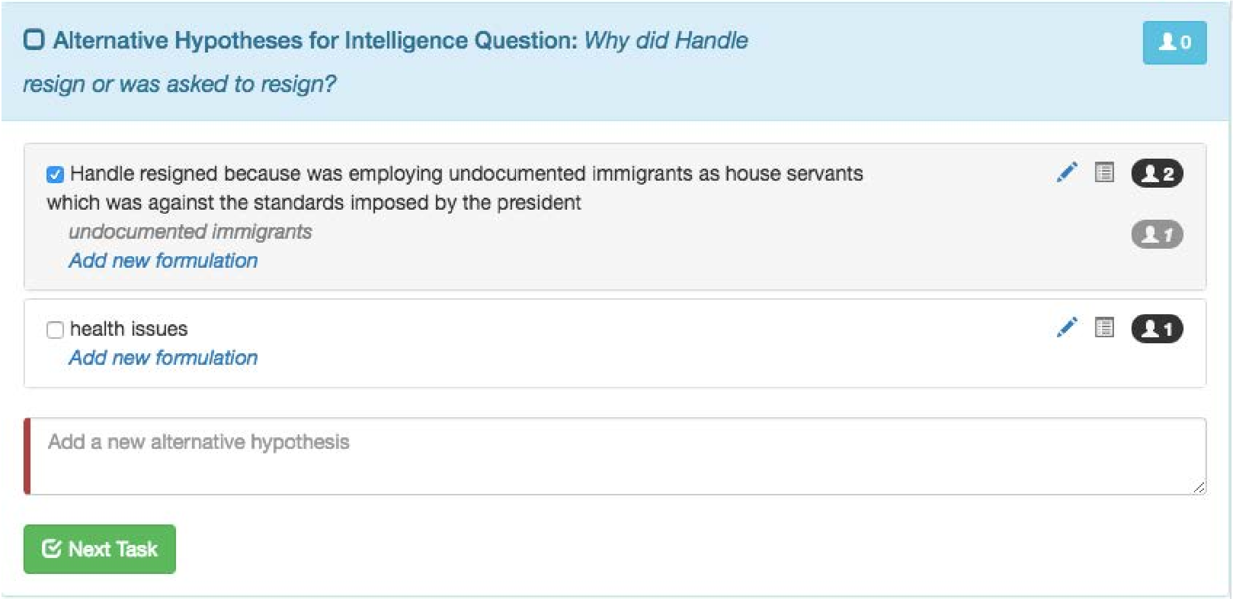
○ undocumented immigrants 1

● Handle resigned because was employing undocumented immigrants as house servants which was against the standards imposed by the president 2

○ Add a new version of alternative hypothesis

Figure 16: Selecting a hypothesis reformulation.

Now Mark's version is supported by more analysts and is shown as the team's version in Figure 17.


□ **Alternative Hypotheses for Intelligence Question:** *Why did Handle resign or was asked to resign?* 0

☑ Handle resigned because was employing undocumented immigrants as house servants which was against the standards imposed by the president 2

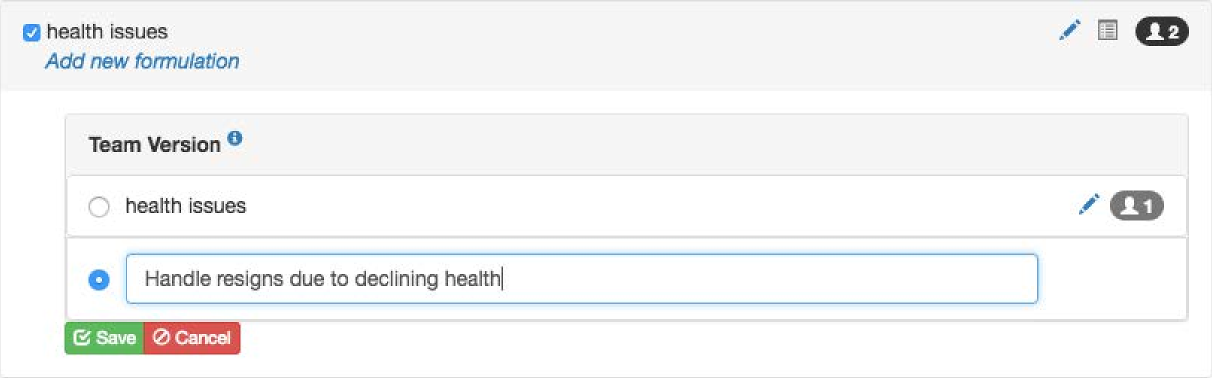
undocumented immigrants 1

□ health issues 1

Add a new alternative hypothesis

Figure 17: Automatic selection of team's solution.

Next, Mary reformulates the second hypothesis, as shown in Figure 18.



☑ health issues

Add new formulation

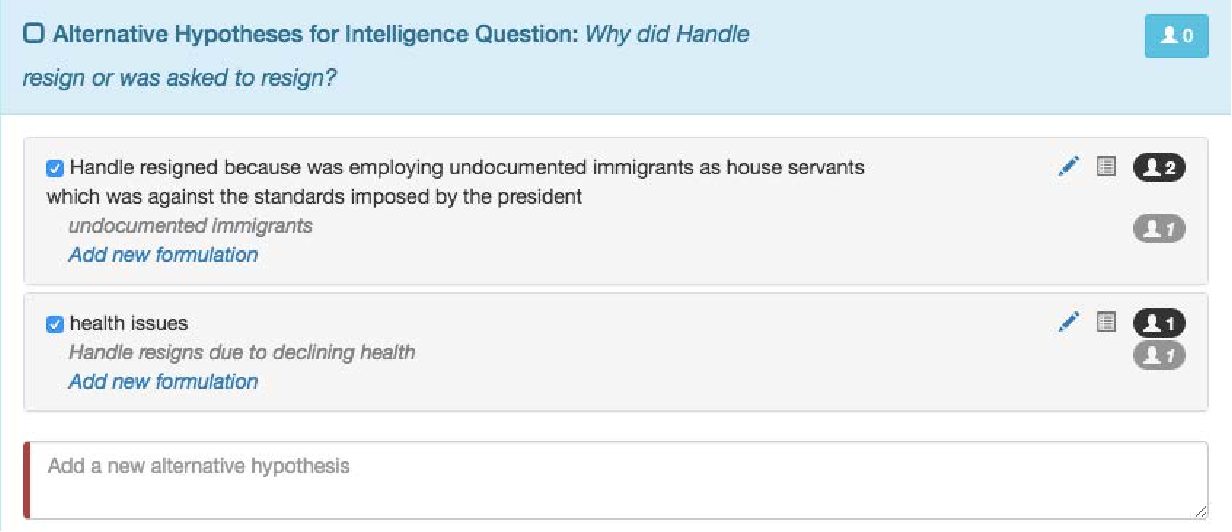
Team Version

○ health issues

● Handle resigns due to declining health

Figure 18: Providing a new reformulation for a hypothesis.

After Mary's reformulation of the second hypothesis, John's version is still the crowd version because each version has 1 vote and John's is older (see Figure 19).



□ Alternative Hypotheses for Intelligence Question: *Why did Handle resign or was asked to resign?*

☑ Handle resigned because was employing undocumented immigrants as house servants which was against the standards imposed by the president

undocumented immigrants

Add new formulation

☑ health issues

Handle resigns due to declining health

Add new formulation

Add a new alternative hypothesis

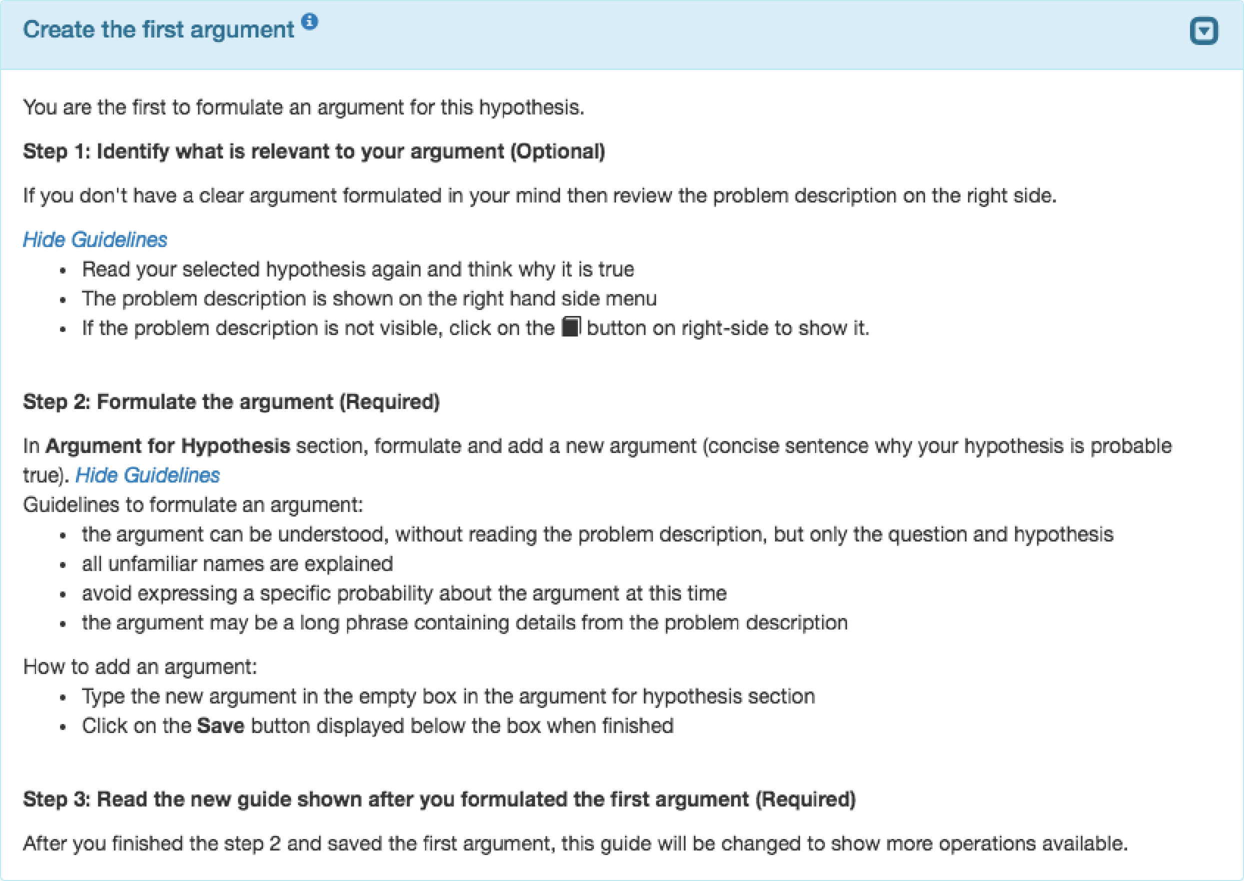
Figure 19: Alternative hypotheses with reformulations.

This process continues, either with other team members joining the brainstorming, or with John or Mark continuing the informal analysis.

3.2.3. Brainstorming Informal Arguments

After brainstorming on the list of alternative hypotheses, the users advance to brainstorm informal arguments for the hypotheses.

Let us imagine that Sara is back and, after agreeing with the alternative hypotheses formulated by John, Mark and Mary, she continues with the informal arguments. Figure 20 shows the guidance offered to the user to define the first informal argument of a hypothesis.



Create the first argument ℹ

You are the first to formulate an argument for this hypothesis.

Step 1: Identify what is relevant to your argument (Optional)

If you don't have a clear argument formulated in your mind then review the problem description on the right side.

Hide Guidelines

- Read your selected hypothesis again and think why it is true
- The problem description is shown on the right hand side menu
- If the problem description is not visible, click on the 📓 button on right-side to show it.

Step 2: Formulate the argument (Required)

In **Argument for Hypothesis** section, formulate and add a new argument (concise sentence why your hypothesis is probable true). *Hide Guidelines*
Guidelines to formulate an argument:

- the argument can be understood, without reading the problem description, but only the question and hypothesis
- all unfamiliar names are explained
- avoid expressing a specific probability about the argument at this time
- the argument may be a long phrase containing details from the problem description

How to add an argument:

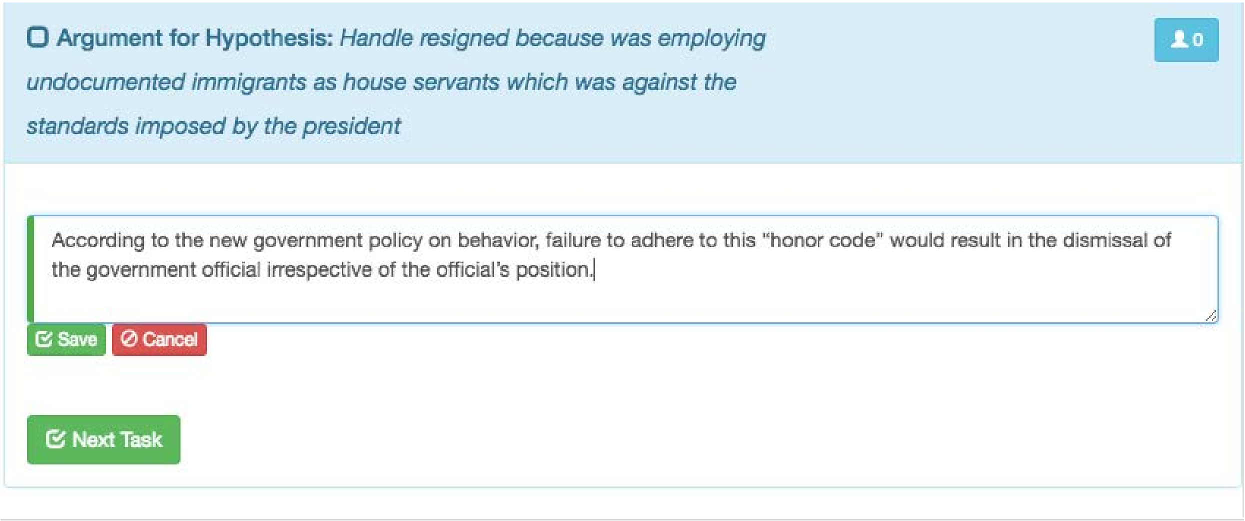
- Type the new argument in the empty box in the argument for hypothesis section
- Click on the **Save** button displayed below the box when finished

Step 3: Read the new guide shown after you formulated the first argument (Required)

After you finished the step 2 and saved the first argument, this guide will be changed to show more operations available.

Figure 20: Guidance to define the first informal argument of a hypothesis.

Sara continues defining informal arguments, as shown in Figure 21. The process of brainstorming informal arguments is similar to the process to brainstorm alternative hypotheses presented in Section 3.2.2.

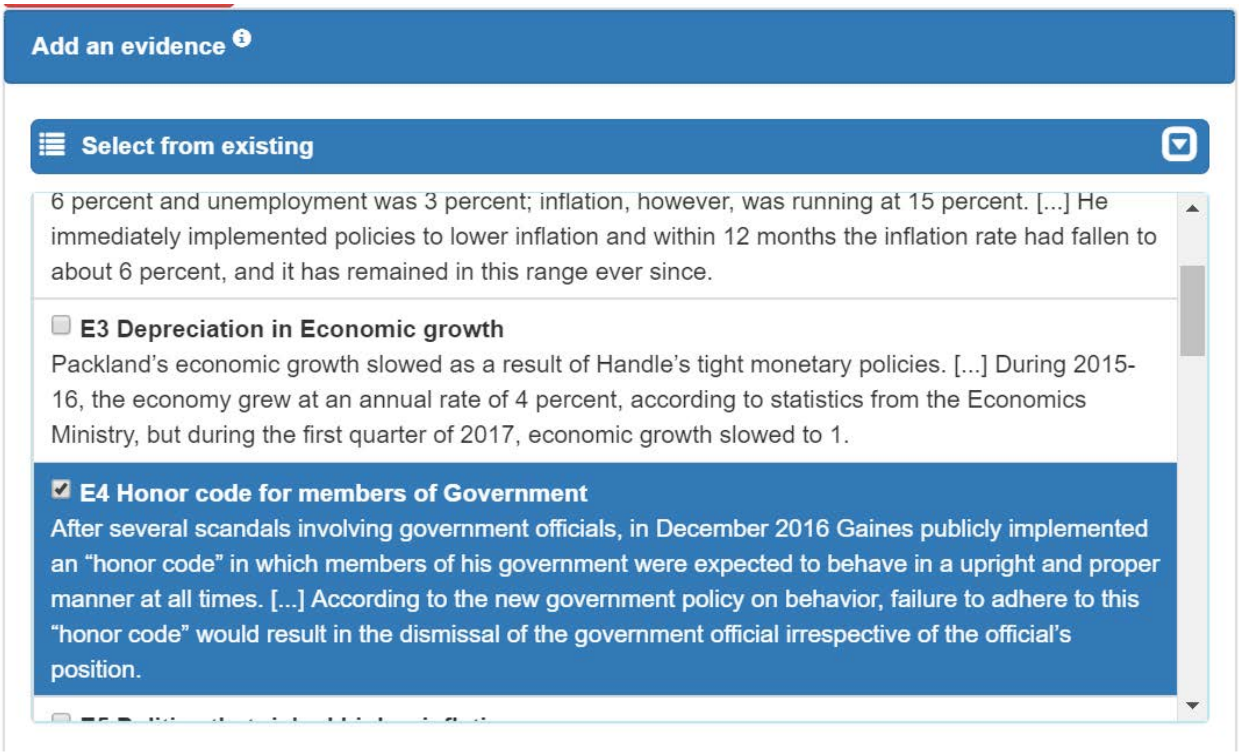

□ **Argument for Hypothesis:** *Handle resigned because was employing undocumented immigrants as house servants which was against the standards imposed by the president*

According to the new government policy on behavior, failure to adhere to this “honor code” would result in the dismissal of the government official irrespective of the official’s position.

Figure 21: Defining an informal argument.

3.2.4. Associating Favoring and Disfavoring Evidence

After defining an informal argument, Sara continues the informal analysis by associating favoring and disfavoring evidence to the informal argument, either from a list of predefined items of evidence (as shown in Figure 22), or by creating new items of evidence based on the problem description.



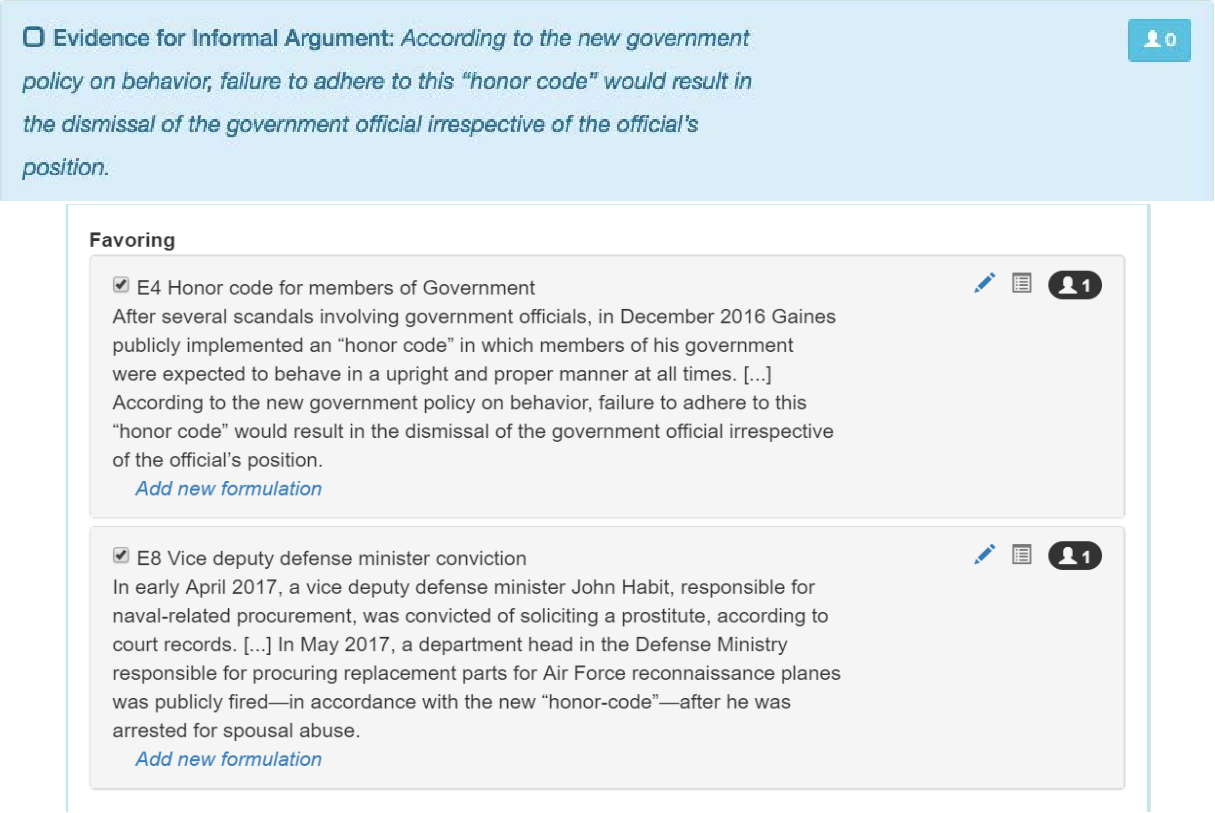
6 percent and unemployment was 3 percent; inflation, however, was running at 15 percent. [...] He immediately implemented policies to lower inflation and within 12 months the inflation rate had fallen to about 6 percent, and it has remained in this range ever since.

□ **E3 Depreciation in Economic growth**
Packland’s economic growth slowed as a result of Handle’s tight monetary policies. [...] During 2015-16, the economy grew at an annual rate of 4 percent, according to statistics from the Economics Ministry, but during the first quarter of 2017, economic growth slowed to 1.

☑ **E4 Honor code for members of Government**
After several scandals involving government officials, in December 2016 Gaines publicly implemented an “honor code” in which members of his government were expected to behave in a upright and proper manner at all times. [...] According to the new government policy on behavior, failure to adhere to this “honor code” would result in the dismissal of the government official irrespective of the official’s position.

Figure 22: Selecting existing evidence.

At the end of the process, all the evidence items selected by the team members for the informal argument, are shown in a single list (see Figure 23).



□ **Evidence for Informal Argument:** *According to the new government policy on behavior, failure to adhere to this "honor code" would result in the dismissal of the government official irrespective of the official's position.*

Favoring

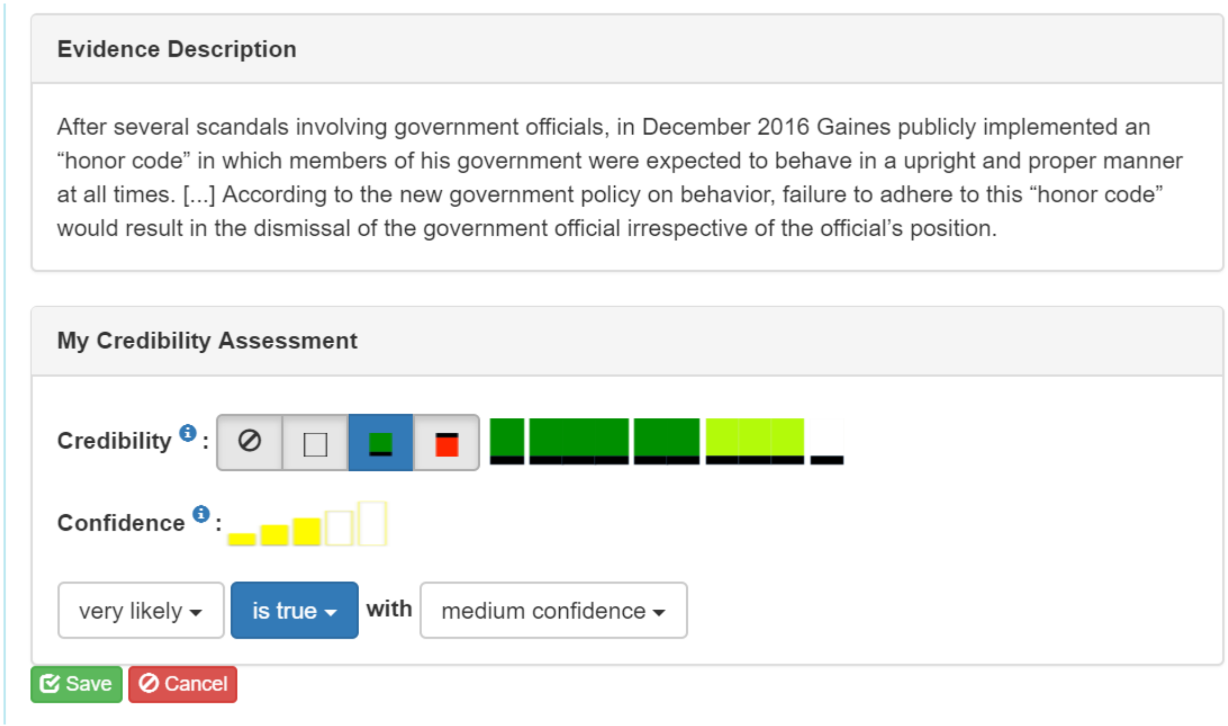
☑ E4 Honor code for members of Government
After several scandals involving government officials, in December 2016 Gaines publicly implemented an "honor code" in which members of his government were expected to behave in a upright and proper manner at all times. [...] According to the new government policy on behavior, failure to adhere to this "honor code" would result in the dismissal of the government official irrespective of the official's position.
Add new formulation

☑ E8 Vice deputy defense minister conviction
In early April 2017, a vice deputy defense minister John Habit, responsible for naval-related procurement, was convicted of soliciting a prostitute, according to court records. [...] In May 2017, a department head in the Defense Ministry responsible for procuring replacement parts for Air Force reconnaissance planes was publicly fired—in accordance with the new "honor-code"—after he was arrested for spousal abuse.
Add new formulation

Figure 23: Selected favoring evidence for an informal argument.

3.2.5. Brainstorming the Credibility of Evidence

Another important task performed by the team is to brainstorm the credibility of evidence. First a team member provides his/her own credibility assessment, and after that the average team credibility assessment is shown, and then updated based on subsequent assessments by other team members. When assessing the credibility of an evidence item, users may also specify their confidence in the assessment made (see Figure 24).



Evidence Description

After several scandals involving government officials, in December 2016 Gaines publicly implemented an "honor code" in which members of his government were expected to behave in a upright and proper manner at all times. [...] According to the new government policy on behavior, failure to adhere to this "honor code" would result in the dismissal of the government official irrespective of the official's position.

My Credibility Assessment

Credibility : ⊘ □ 🟩 🟥

Confidence :

very likely ▾ is true ▾ with medium confidence ▾

Figure 24: Assessing the credibility of evidence.

3.2.6. Another Illustration of Informal Analysis Collaboration

In the following illustration we consider the Cesium-137 problem from Table 2: It was reported that a canister containing cesium-137 is missing from the XYZ Company in MD. The question is: What happened to it?

Because the question is provided, the first task for the participants is to find potential answers or hypotheses. In this illustration we will assume a team of 3 participants, P1, P2 and P3, with P1 being the first to formulate the answers:

What happened to the cesium-137 canister?

- Was stolen
- Was misplaced
- Was lost

Next P1 continues with the formulation of informal arguments to support or refute each of these hypotheses. However, in this example, we focus on the contributions of the other two participants to the definition of alternative hypotheses.

Next to work on the analysis is P2. He reads the answers provided by P1 and has the following options:

- Reformulate an existing answer;
- Vote for the formulation of an existing answer;
- Provide a new answer;
- Reject an existing answer.

Situation:
Today's Washington Gazette published an article about how securely radioactive materials are stored at facilities in the DC area. Willard, the investigative reporter and author of this piece notes his discovery that a canister containing cesium-137 is missing from the XYZ Company in MD, just three days ago. The XYZ Company manufactures devices for sterilizing medical equipment and uses cesium-137 in these devices along with other radioactive materials. This piece arouses your curiosity because of your concern about terrorists planting dirty bombs in our cities.
Question: What happened to the cesium-137 canister?
Available Information:
Contacted about the cesium-137 canister, Ralph, the supervisor of the warehouse, reports that the cesium-137 canister is registered as being in the warehouse, that no one at the XYZ company had checked it out, but it is not located anywhere in the hazardous materials locker. He also indicates that the lock on the hazardous materials locker appears to have been forced.

The professional locksmith Clyde was consulted, and he said that the lock had been forced open, but it was a clumsy job.

The security camera of the XYZ warehouse contains a video segment showing a person loading a canister into a U-Haul panel truck at 2:45PM. The video camera stopped functioning at 3:00PM that day and was not repaired until later that evening. The video camera did not observe any other canisters being removed that day prior to 2:45PM.

There is a security perimeter around the XYZ warehouse and employee parking area that has just one gate that is controlled by a guard. On the day before the canister was discovered missing, the security guard, Sam, recorded that a panel truck having Maryland license plate MDC-578 was granted entry at 2:15PM. The driver of this vehicle showed the guard a manifest containing items being delivered to the XYZ warehouse. This manifest contained a list of packing materials allegedly ordered by the XYZ Company. The vehicle was allowed to enter the parking area. Deliveries are only accepted between the hours of 7AM and 5PM.

Grace, the Vice President for Operations at XYZ, tells us that while they have several projects involving hazardous materials, none uses cesium-137.

Every week security personnel conduct an inventory on all canisters containing hazardous material. The week before the canister was discovered missing, the cesium canister was noted in company records as being in its assigned location.

Security personnel at XYZ are rigorously vetted for honesty and trustworthiness.

Table 2: The "Cesium 137" problem.

P2 decides to reformulate two of the answers. He also rejects the third answer, justifies the rejection, and provides another answer. The view of the informal analysis for P2 is:

What happened to the cesium-137 canister?

- P2: The cesium-137 canister was stolen (1 vote)
  Team version: Was stolen (1 vote)
- P2: The cesium-137 canister was misplaced (1 vote)
  Team version: Was misplaced (1 vote)
- P2: The cesium-137 canister is being used in a project of the XYZ Company without having been checked out from XYZ warehouse (1 vote)

After that P2 reviews and revises the analysis done by P1 for the first two hypotheses and provides an analysis for the new hypothesis proposed by him.

P3 reads the hypotheses proposed by P1 and P2 and votes for the reformulations proposed by P2. After reading the note from P2, P3 agrees that the "Was lost" hypothesis is covered by the proposed answers.

The team result is the following one:

What happened to the cesium-137 canister?

- Team version: The cesium-137 canister was stolen (2 votes: P2, P3)
  P1 version: Was stolen (1 vote: P1)
- Team version: The cesium-137 canister was misplaced (2 votes: P2, P3)
  P1 version: Was misplaced (1 vote: P1)
- Team version: Was lost (1 vote: P1)
- Team version: The cesium-137 canister is being used in a project of the XYZ Company without having been checked out from XYZ warehouse (2 votes: P2, P3)

Because P2 and P3 have modified the hypotheses, they are marked as incomplete for P1. Therefore, the next time P1 logs in, the system guides her to review the modifications proposed by P2 and P3. Let us assume that P1 also votes for these modifications. As a result, the initial formulations of the hypotheses remain without any vote and are deleted.

The same process is used for the follow-on brainstorming task of providing informal arguments for each of the hypotheses. For example, the informal argument developed by the team for the hypothesis "The cesium-137 canister was stolen," is the following one:

> A truck entered the company, the canister was stolen from the locker, loaded into the truck, and the truck left with the canister.

After brainstorming informal arguments, the participants associate favoring and disfavoring evidence to each of these arguments.

The last brainstorming phase is to assess the credibility of the evidence used. Each participant is asked to assess the credibility of each item of evidence, and these individual assessments are combined into a team assessment. The final informal argumentation for "The cesium-137 canister was stolen" is shown in Figure 25. For example, E1 was assessed as likely (55-70%) by P1, as more than likely (70-80%) by P2, and as barely likely (50-55%) by P3, resulting in a team credibility of likely (55-70%).

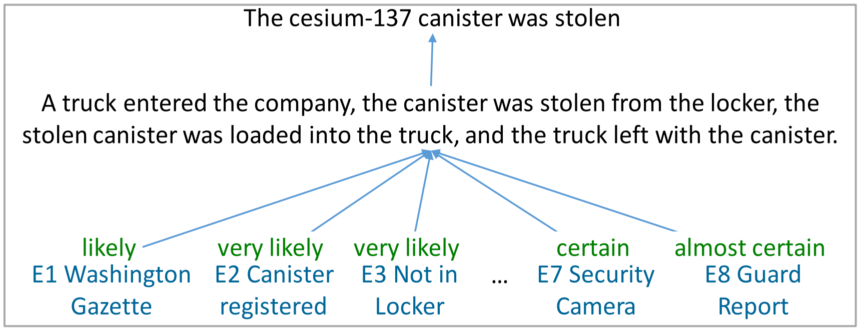

The cesium-137 canister was stolen
A truck entered the company, the canister was stolen from the locker, the stolen canister was loaded into the truck, and the truck left with the canister.
likely
E1 Washington Gazette
very likely
E2 Canister registered
very likely
E3 Not in Locker
…
certain
E7 Security Camera
almost certain
E8 Guard Report

Figure 25: Informal argumentation.

To summarize, the participants first brainstorm the reformulation of each intelligence question, and then the formulation of its potential answers (i.e., hypotheses). After that they brainstorm informal arguments for each hypothesis, relevant evidence for each argument, and credibility for each evidence item. They have two means to perform the brainstorming: ***problem check-list*** (a guided step by step process, based on a list of tasks that need to be completed in a predefined order), and ***graphical analysis*** (a graphical way providing the freedom to perform the tasks in any possible order).

3.3. Development of the Formal Analysis with Cogent

3.3.1. Cogent Interface

Figure 26 illustrates the interface of Cogent. The upper left side is the Whiteboard area where the argumentation is developed using simple right-click and drag-and-drop operations. The upper right side is the assistants' area, each assistant helping in performing a group of related operations. The bottom part is the help area.

The Whiteboard displays the argumentation graphically in a tree-like structure and provides operations for modifying and extending it. The user can hover the mouse cursor over an element in the argumentation to view more details about it (when available), click on an element to select it, double-click on it to modify it (if possible), or right-click on the element to select from a list of actions applicable to that element. For example, double-clicking on a hypothesis opens a floating editor that allows the user to modify the description of that hypothesis. The new description is saved when the user clicks outside of the editor window. Double-clicking on a probability icon (relevance or credibility) displays a list of values from which the user can select a new value for that element.

Right-clicking on a hypothesis displays a list of actions applicable to it, such as adding a favoring argument under it, adding a sibling to it in an "AND" argument, or deleting it (and everything under it) from the argumentation.

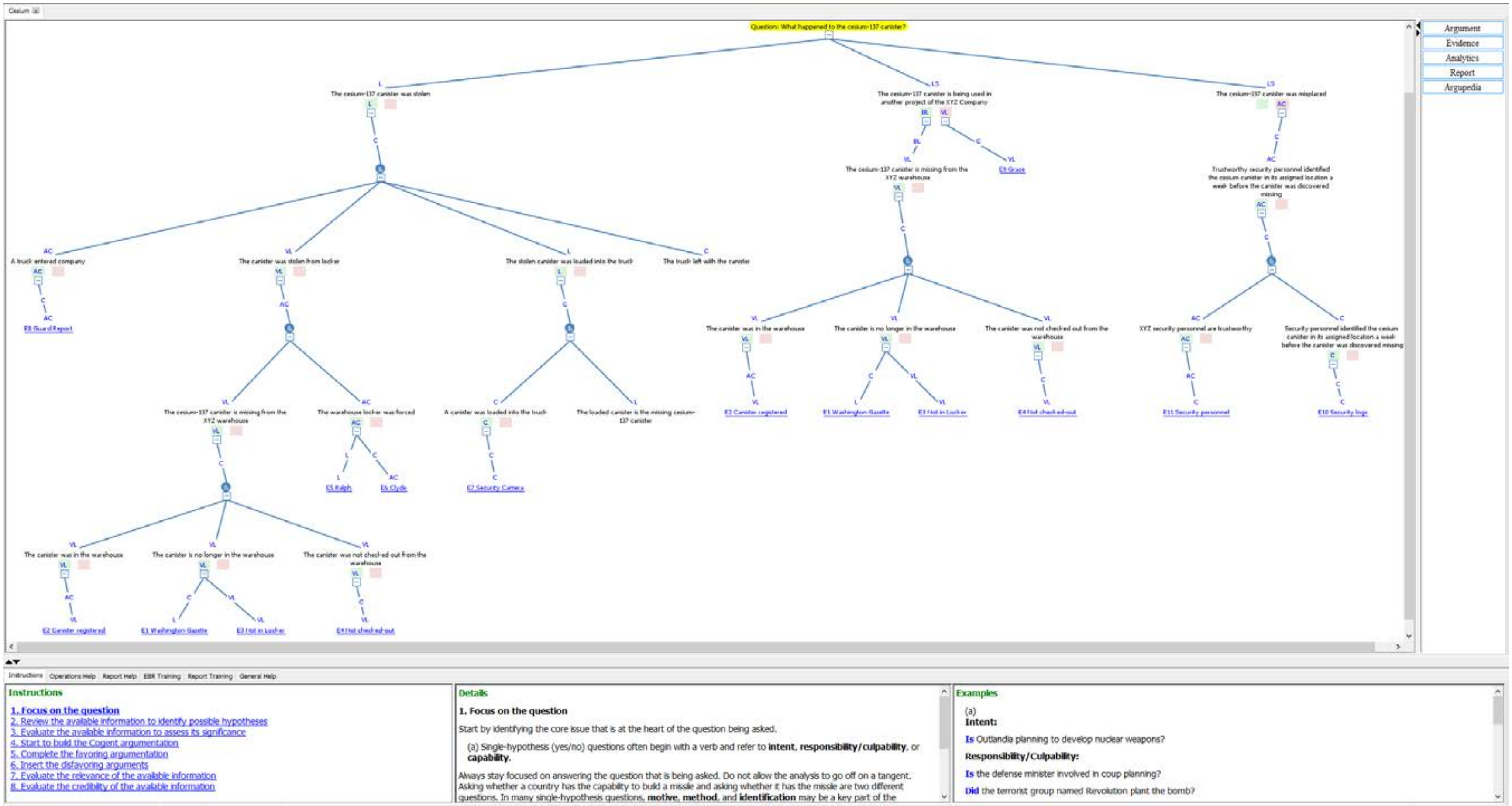
Selecting an element from the argumentation allows other Cogent tools and assistants to display additional information about it. For example, selecting an item of evidence in the argumentation by clicking on it opens it in the Evidence Assistant, allowing the user to see additional information associated with it (such as a reference link) and to modify it, while the Help tool (shown at the bottom in Figure 26) displays information about the actions that can be performed on the evidence item in the Whiteboard (in the Operations Help tab).

The Whiteboard also supports drag-and-drop to perform several types of operations. For example, the user can copy or move sub-arguments from one place of the argumentation to another, and also associate evidence with a hypothesis by dragging it from the Evidence Assistant and dropping it on the favoring (green) or disfavoring (red) squares below the hypothesis.

Basic drag-and-drop is also allowed from external tools. For example, the user can select some text in a Word document, drag it over to the Whiteboard, and drop it over a hypothesis – as a result Cogent will create a new item of evidence with the selected text as its description, will automatically open it in the Evidence Assistant for further refinement, and will associate it with the hypothesis under which it was dropped.

3.3.2. Argument Development

The informal analysis developed in Argupedia is imported into Cogent by each member of the team who continues to work independently to develop his or her own formal analysis. For example, guided by the informal argumentation in Figure 25, a user has developed the formal argumentation from Figure 27. This is generally done by decomposing the top hypothesis into simpler and simpler hypotheses, down to the level of simple hypotheses that are assessed based on the relevant evidence. At this point, disfavoring arguments and evidence may also be inserted into the developed argumentation.


Cesium
Question: What happened to the cesium-137 canister?
The cesium-137 canister was stolen
The cesium-137 canister is being used in another project of the XYZ Company
The cesium-137 canister was misplaced
A truck entered company
The canister was stolen from locker
The stolen canister was loaded into the truck
The truck left with the canister
E8 Guard Report
The cesium-137 canister is missing from the XYZ warehouse
The warehouse locker was forced
A canister was loaded into the truck
The loaded canister is the missing cesium-137 canister
E5 Ralph
E6 Clyde
E7 Security Camera
The canister was in the warehouse
The canister is no longer in the warehouse
The canister was not checked out from the warehouse
E2 Canister registered
E1 Washington Gazette
E3 Not in Locker
E4 Not checked-out
E9 Grace
Trustworthy security personnel identified the cesium canister in its assigned location a week before the canister was discovered missing
XYZ security personnel are trustworthy
Security personnel identified the cesium canister in its assigned location a week before the canister was discovered missing
E11 Security personnel
E10 Security logs
Argument
Evidence
Analytics
Report
Argupedia
Instructions
Operations Help
Report Help
EBR Training
Report Training
General Help

Instructions

1. Focus on the question
2. Review the available information to identify possible hypotheses
3. Evaluate the available information to assess its significance
4. Start to build the Cogent argumentation
5. Complete the favoring argumentation
6. Insert the disfavoring arguments
7. Evaluate the relevance of the available information
8. Evaluate the credibility of the available information

Details

1. Focus on the question

Start by identifying the core issue that is at the heart of the question being asked.

(a) Single-hypothesis (yes/no) questions often begin with a verb and refer to **intent**, **responsibility/culpability**, or **capability**.

Always stay focused on answering the question that is being asked. Do not allow the analysis to go off on a tangent. Asking whether a country has the capability to build a missile and asking whether it has the missile are two different questions. In many single-hypothesis questions, **motive**, **method**, and **identification** may be a key part of the

Examples

(a)
Intent:

Is Outlandia planning to develop nuclear weapons?

Responsibility/Culpability:

Is the defense minister involved in coup planning?

Did the terrorist group named Revolution plant the bomb?

Figure 26: Cogent interface.

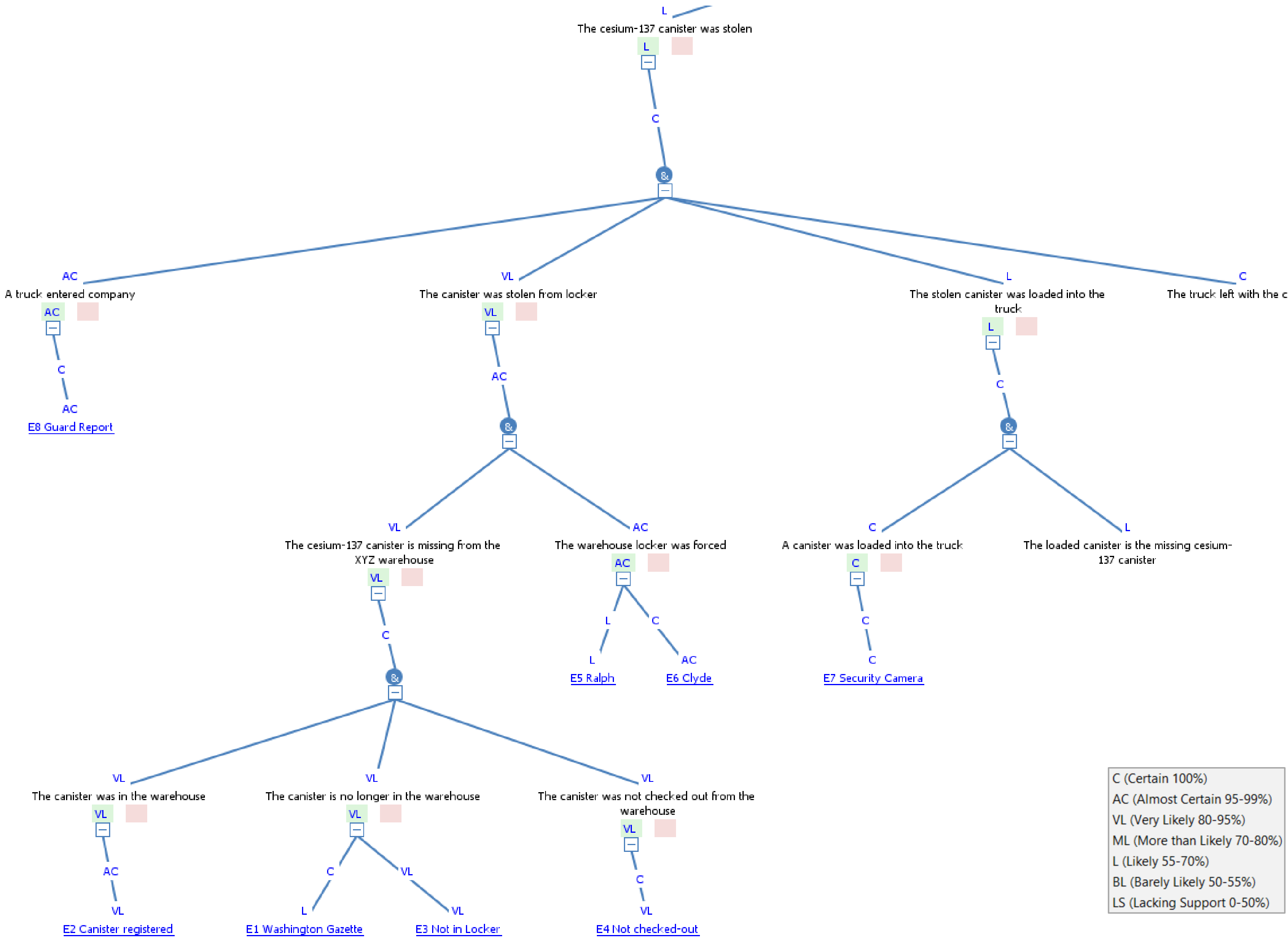

L
The cesium-137 canister was stolen
L
C
&
AC
A truck entered company
AC
C
AC
E8 Guard Report
VL
The canister was stolen from locker
VL
AC
&
L
The stolen canister was loaded into the truck
L
C
&
C
The truck left with the canister
VL
The cesium-137 canister is missing from the XYZ warehouse
VL
C
&
AC
The warehouse locker was forced
AC
L
C
L
E5 Ralph
AC
E6 Clyde
C
A canister was loaded into the truck
C
C
C
E7 Security Camera
L
The loaded canister is the missing cesium-137 canister
VL
The canister was in the warehouse
VL
AC
VL
E2 Canister registered
VL
The canister is no longer in the warehouse
VL
C
L
E1 Washington Gazette
VL
VL
E3 Not in Locker
VL
The canister was not checked out from the warehouse
VL
C
VL
E4 Not checked-out
C (Certain 100%)
AC (Almost Certain 95-99%)
VL (Very Likely 80-95%)
ML (More than Likely 70-80%)
L (Likely 55-70%)
BL (Barely Likely 50-55%)
LS (Lacking Support 0-50%)

Figure 27: Formal argumentation.

Figure 28 shows the formal analysis with the three alternative hypotheses and the top parts of their argumentations.

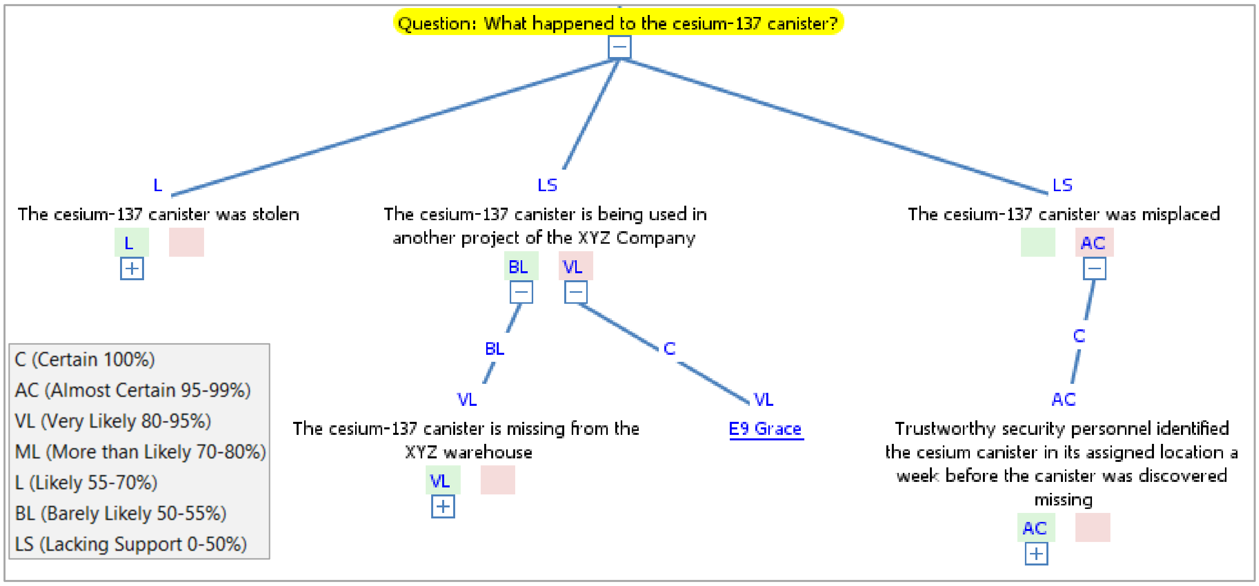
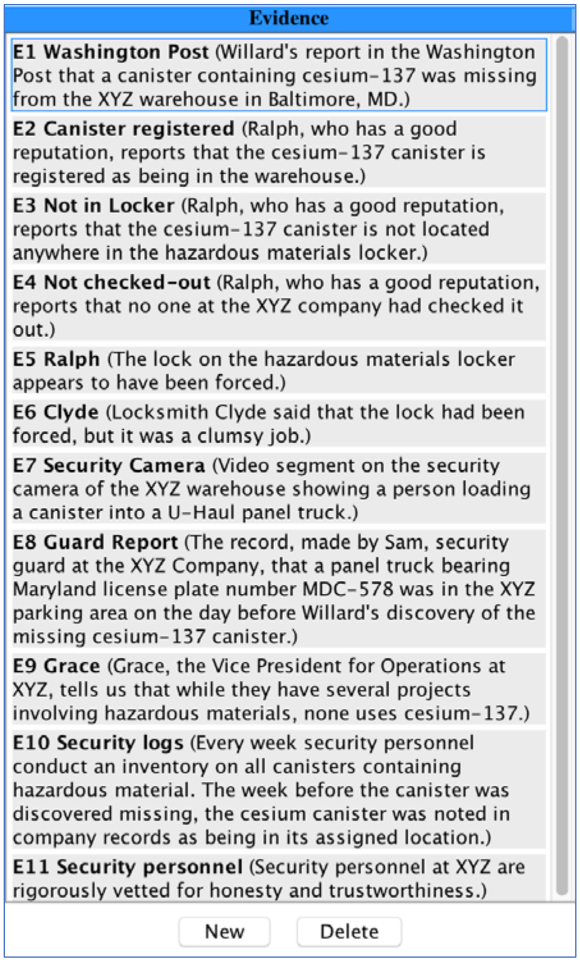

Question: What happened to the cesium-137 canister?
L
The cesium-137 canister was stolen
L
LS
The cesium-137 canister is being used in another project of the XYZ Company
BL
VL
LS
The cesium-137 canister was misplaced
AC
BL
VL
The cesium-137 canister is missing from the XYZ warehouse
VL
C
VL
E9 Grace
C
AC
Trustworthy security personnel identified the cesium canister in its assigned location a week before the canister was discovered missing
AC
C (Certain 100%)
AC (Almost Certain 95-99%)
VL (Very Likely 80-95%)
ML (More than Likely 70-80%)
L (Likely 55-70%)
BL (Barely Likely 50-55%)
LS (Lacking Support 0-50%)

Figure 28: Formal analysis.

3.3.3. Evidence Assistant

The Evidence Assistant allows the user to define, modify, and delete items of evidence. The main interface of the assistant, shown in Figure 29, displays the list of currently available items of evidence. For each item of evidence, the assistant displays its name followed by its description in parentheses. The user can click on an item of evidence to select it, or double-click it to modify it in the editor. Clicking on the “Delete” button below the list deletes the currently selected evidence item, while clicking on “New” opens the editor to define a new item of evidence. The user can drag items of evidence from this list to the Whiteboard to associate them with hypotheses from the argumentation.



Evidence

E1 Washington Post (Willard's report in the Washington Post that a canister containing cesium-137 was missing from the XYZ warehouse in Baltimore, MD.)

E2 Canister registered (Ralph, who has a good reputation, reports that the cesium-137 canister is registered as being in the warehouse.)

E3 Not in Locker (Ralph, who has a good reputation, reports that the cesium-137 canister is not located anywhere in the hazardous materials locker.)

E4 Not checked-out (Ralph, who has a good reputation, reports that no one at the XYZ company had checked it out.)

E5 Ralph (The lock on the hazardous materials locker appears to have been forced.)

E6 Clyde (Locksmith Clyde said that the lock had been forced, but it was a clumsy job.)

E7 Security Camera (Video segment on the security camera of the XYZ warehouse showing a person loading a canister into a U-Haul panel truck.)

E8 Guard Report (The record, made by Sam, security guard at the XYZ Company, that a panel truck bearing Maryland license plate number MDC-578 was in the XYZ parking area on the day before Willard's discovery of the missing cesium-137 canister.)

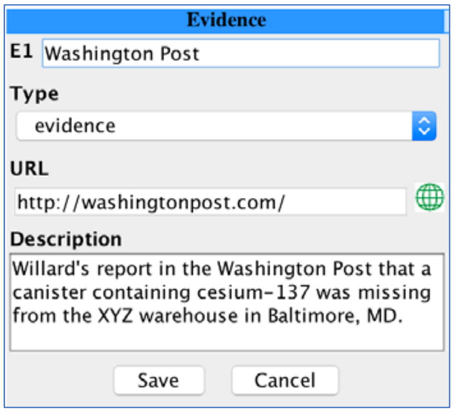
E9 Grace (Grace, the Vice President for Operations at XYZ, tells us that while they have several projects involving hazardous materials, none uses cesium-137.)

E10 Security logs (Every week security personnel conduct an inventory on all canisters containing hazardous material. The week before the canister was discovered missing, the cesium canister was noted in company records as being in its assigned location.)

E11 Security personnel (Security personnel at XYZ are rigorously vetted for honesty and trustworthiness.)

Figure 29: Evidence Assistant interface.

The Evidence editor interface is shown in Figure 30. The user can modify the name of the evidence item, its type and its description, and can provide a URL to the source. Clicking on the “Save” button saves the modifications made by the user, while clicking on “Cancel” discards all of them. In both cases the editor is closed and the Evidence Assistant re-displays the list of available items of evidence.



Evidence

E1 Washington Post

Type

evidence

URL

http://washingtonpost.com/

Description

Willard's report in the Washington Post that a canister containing cesium-137 was missing from the XYZ warehouse in Baltimore, MD.

Figure 30: Evidence editor interface.

3.3.4. Analytics Assistant

The Analytics Assistant allows the user to check the current argumentation for biases and other errors and warnings.

3.3.4.1. Bias Avoidance, Detection, and Mitigation

A wide variety of biases can reduce the accuracy of intelligence analysis. There are several complementary ways by which Cogent helps users recognize and counter many of the biases.

With Cogent the user performs a rigorous evidence-based hypothesis analysis. This analysis shows and justifies all the reasoning steps, evidence, assumptions, and probabilistic assessments. This explicit analysis, by itself, counters several biases, such as the *anchoring bias* (the tendency to rely too heavily, or "anchor", on one trait or piece of information), the *vividness bias* (the tendency to overweight vivid or prestigious attributes and underweight less flashy issues), the *availability bias* (a mental shortcut that relies on immediate examples that come to a person's mind), the *persistence of impressions based on discredited evidence* (the tendency to still believe evidence that was discredited), and the *conjunction fallacy* (assuming that a conjunction of conditions is more probable than one of them).

Currently, the Analytics Assistant of Cogent includes methods to automatically detect the following biases in an argumentation:

The *confirmation bias* (the tendency to seek only that information that is consistent with the preferred hypothesis) is signaled when a hypothesis has only favoring arguments and evidence.

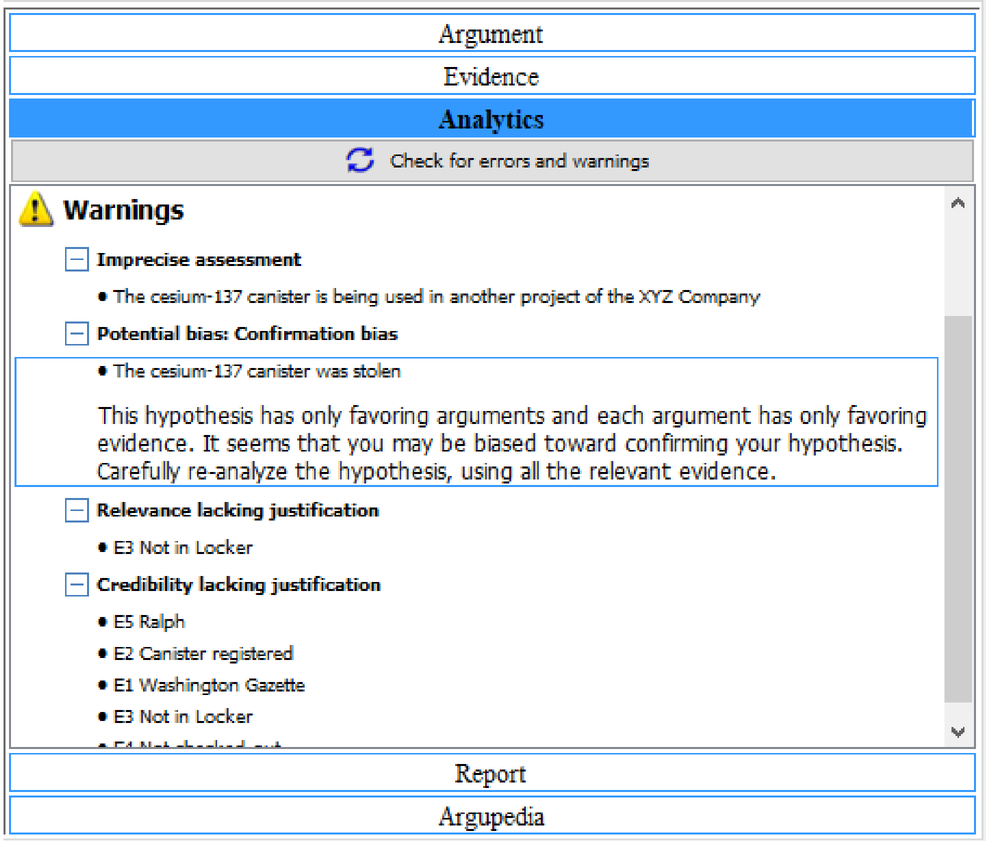
The *satisficing bias* (choosing the first hypothesis that appears good enough rather than carefully identifying all possible hypotheses and determining which one is the most consistent with the evidence) is signaled when the user has analyzed only one of the possible hypotheses, ignoring its alternatives. It is also signaled when several hypotheses are analyzed, but one of them has a significantly larger argumentation, with a number of nodes more than 5 times larger than each of the other ones. Of course, 5 is an adjustable parameter.

The *absence of evidence bias* (failure to consider the degree of completeness of the available evidence) is signaled when for each top argument of each hypothesis there are more assumptions than evidence.

For each identified bias, Cogent explains it and advises the user on how to mitigate it.

3.3.4.2. Automatic Detection of Errors and Warnings

The assistant computes an initial list of errors and warnings the first time its interface is opened for an analysis, groups them based on category and type, and displays them to the user, as shown in Figure 31. If the user hovers the mouse over an error or warning item, the assistant provides a brief description of the action to take in a tooltip. If the user clicks on the item, the assistant expands it and shows a more detailed description of how to fix the issue along with selecting the corresponding argumentation fragment in the Whiteboard. Notice in Figure 31 that the user is alerted about the potential *bias toward confirming* the hypothesis “The cesium-137 canister was stolen” because this hypothesis has only favoring arguments and each such argument has only favoring evidence.



Warnings

Imprecise assessment

- The cesium-137 canister is being used in another project of the XYZ Company

Potential bias: Confirmation bias

- The cesium-137 canister was stolen

This hypothesis has only favoring arguments and each argument has only favoring evidence. It seems that you may be biased toward confirming your hypothesis. Carefully re-analyze the hypothesis, using all the relevant evidence.

Relevance lacking justification

- E3 Not in Locker

Credibility lacking justification

- E5 Ralph
- E2 Canister registered
- E1 Washington Gazette
- E3 Not in Locker
- E4 Not checked out

Figure 31: Messages from the Analytics Assistant.

Because checking the argumentation for errors and warnings can take some time, the assistant will not automatically update the list of identified issues after each modification performed by the user in the Whiteboard. Instead, the user can click on the “Check for errors and warnings” button at the top of the Analytics Assistant to request an updated list of issues.

The current version of Cogent provides warnings for the following weaknesses in the argumentation, together with guidance on resolving them:

- evidence lacking description;
- assumption lacking justification;
- relevance lacking justification;
- credibility lacking justification; and
- incompletely defined item (it could be an evidence item, a question, or a hypothesis).

Co-Arg also detects the following errors with appropriate correction guidance:

- sum of probabilities of top hypotheses < 100%;
- sum of probabilities of top hypotheses > 100% (for disjoint hypotheses);
- credibility not assessed;
- probability not assessed;
- relevance not assessed;

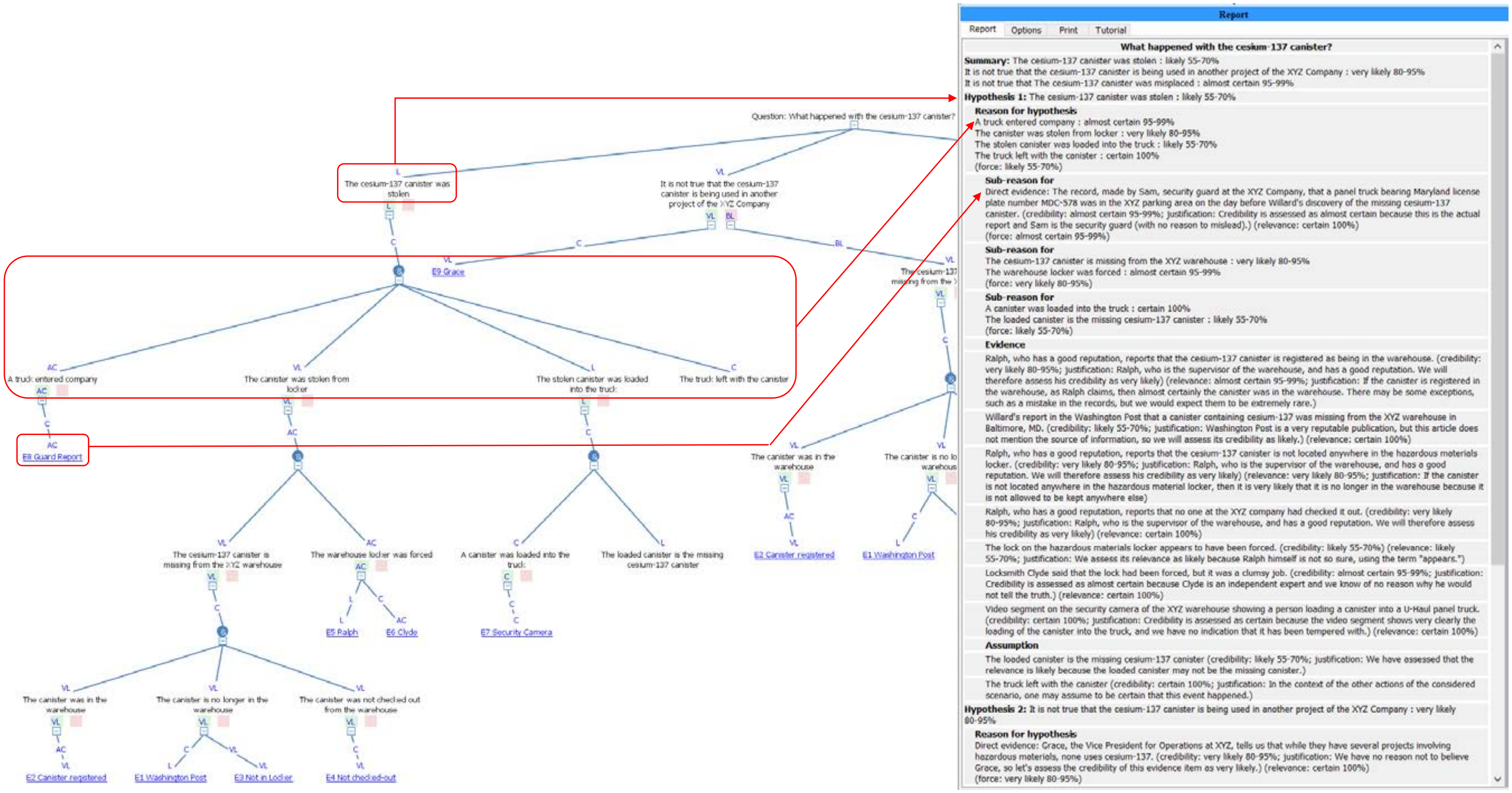
- relevance lacking support;
- inconsistent relevance of arguments;
- duplicated evidence;
- duplicated evidence and inconsistent relevance;
- duplicated hypothesis;
- duplicated sub-hypothesis and inconsistent relevance;
- duplicated “AND” argument;
- duplicated “AND” argument and inconsistent relevance.

.

3.3.5. Report Assistant

The Report Assistant component of Cogent facilitates the creation of a high-quality analytic product that not only answers the intelligence questions but also describes the reasoning and presents it in a way that is easy to understand and easy to navigate when more details are needed. The assistant is invoked once the argumentation is finalized and checked for errors, warnings and biases. At this point, the user can open the assistant and click on a button to generate a report based on the argumentation. This automatically generated report takes the argumentation, including the probabilities of the hypotheses and the most important parts of the reasoning tree (upper level hypotheses, evidence and assumptions) and converts it to a structured textual report that verbalizes the top levels of the argumentation and adds the corresponding evidence and assumptions, as illustrated in Figure 32.

The structured report can be further edited into a production report, to make it more understandable and persuasive. This includes editing the generated sentences, inserting argument fragments, and adding the corresponding evidence. Figure 33 shows an example of a production report corresponding to the structured report from Figure 32. It consists of a textual description that includes a probability for each hypothesis and its reasoning followed by a more detailed appendix. The textual description includes links to the inserted argument fragments and evidence that are part of the appendix.


Question: What happened with the cesium-137 canister?
The cesium-137 canister was stolen
It is not true that the cesium-137 canister is being used in another project of the XYZ Company
E9 Grace
The cesium-137 missing from the
A truck entered company
The canister was stolen from locker
The stolen canister was loaded into the truck
The truck left with the canister
E8 Guard Report
The cesium-137 canister is missing from the XYZ warehouse
The warehouse locker was forced
A canister was loaded into the truck
The loaded canister is the missing cesium-137 canister
The canister was in the warehouse
The canister is no lo warehous
E2 Canister registered
E1 Washington Post
E5 Ralph
E6 Clyde
E7 Security Camera
The canister was in the warehouse
The canister is no longer in the warehouse
The canister was not checked out from the warehouse
E2 Canister registered
E1 Washington Post
E3 Not in Locker
E4 Not checked-out
Report
Report
Options
Print
Tutorial

What happened with the cesium-137 canister?

Summary: The cesium-137 canister was stolen : likely 55-70%
It is not true that the cesium-137 canister is being used in another project of the XYZ Company : very likely 80-95%
It is not true that The cesium-137 canister was misplaced : almost certain 95-99%

Hypothesis 1: The cesium-137 canister was stolen : likely 55-70%

Reason for hypothesis
A truck entered company : almost certain 95-99%
The canister was stolen from locker : very likely 80-95%
The stolen canister was loaded into the truck : likely 55-70%
The truck left with the canister : certain 100%
(force: likely 55-70%)

Sub-reason for
Direct evidence: The record, made by Sam, security guard at the XYZ Company, that a panel truck bearing Maryland license plate number MDC-578 was in the XYZ parking area on the day before Willard's discovery of the missing cesium-137 canister. (credibility: almost certain 95-99%; justification: Credibility is assessed as almost certain because this is the actual report and Sam is the security guard (with no reason to mislead).) (relevance: certain 100%)
(force: almost certain 95-99%)

Sub-reason for
The cesium-137 canister is missing from the XYZ warehouse : very likely 80-95%
The warehouse locker was forced : almost certain 95-99%
(force: very likely 80-95%)

Sub-reason for
A canister was loaded into the truck : certain 100%
The loaded canister is the missing cesium-137 canister : likely 55-70%
(force: likely 55-70%)

Evidence

Ralph, who has a good reputation, reports that the cesium-137 canister is registered as being in the warehouse. (credibility: very likely 80-95%; justification: Ralph, who is the supervisor of the warehouse, and has a good reputation. We will therefore assess his credibility as very likely) (relevance: almost certain 95-99%; justification: If the canister is registered in the warehouse, as Ralph claims, then almost certainly the canister was in the warehouse. There may be some exceptions, such as a mistake in the records, but we would expect them to be extremely rare.)

Willard's report in the Washington Post that a canister containing cesium-137 was missing from the XYZ warehouse in Baltimore, MD. (credibility: likely 55-70%; justification: Washington Post is a very reputable publication, but this article does not mention the source of information, so we will assess its credibility as likely.) (relevance: certain 100%)

Ralph, who has a good reputation, reports that the cesium-137 canister is not located anywhere in the hazardous materials locker. (credibility: very likely 80-95%; justification: Ralph, who is the supervisor of the warehouse, and has a good reputation. We will therefore assess his credibility as very likely) (relevance: very likely 80-95%; justification: If the canister is not located anywhere in the hazardous material locker, then it is very likely that it is no longer in the warehouse because it is not allowed to be kept anywhere else)

Ralph, who has a good reputation, reports that no one at the XYZ company had checked it out. (credibility: very likely 80-95%; justification: Ralph, who is the supervisor of the warehouse, and has a good reputation. We will therefore assess his credibility as very likely) (relevance: certain 100%)

The lock on the hazardous materials locker appears to have been forced. (credibility: likely 55-70%) (relevance: likely 55-70%; justification: We assess its relevance as likely because Ralph himself is not so sure, using the term "appears.")

Locksmith Clyde said that the lock had been forced, but it was a clumsy job. (credibility: almost certain 95-99%; justification: Credibility is assessed as almost certain because Clyde is an independent expert and we know of no reason why he would not tell the truth.) (relevance: certain 100%)

Video segment on the security camera of the XYZ warehouse showing a person loading a canister into a U-Haul panel truck. (credibility: certain 100%; justification: Credibility is assessed as certain because the video segment shows very clearly the loading of the canister into the truck, and we have no indication that it has been tempered with.) (relevance: certain 100%)

Assumption

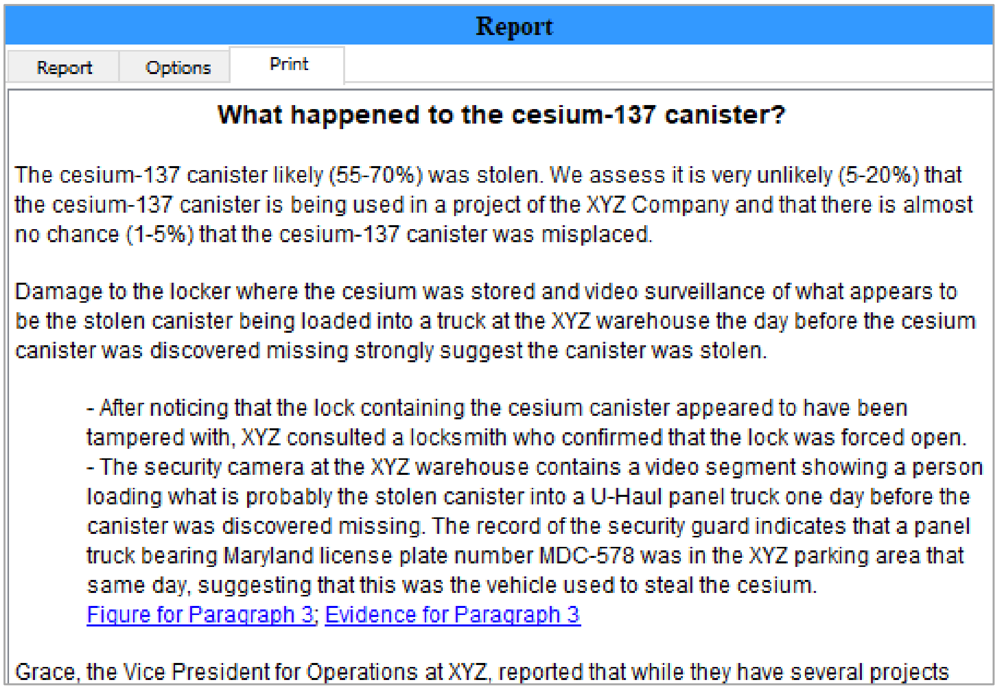
The loaded canister is the missing cesium-137 canister (credibility: likely 55-70%; justification: We have assessed that the relevance is likely because the loaded canister may not be the missing canister.)

The truck left with the canister (credibility: certain 100%; justification: In the context of the other actions of the considered scenario, one may assume to be certain that this event happened.)

Hypothesis 2: It is not true that the cesium-137 canister is being used in another project of the XYZ Company : very likely 80-95%

Reason for hypothesis
Direct evidence: Grace, the Vice President for Operations at XYZ, tells us that while they have several projects involving hazardous materials, none uses cesium-137. (credibility: very likely 80-95%; justification: We have no reason not to believe Grace, so let's assess the credibility of this evidence item as very likely.) (relevance: certain 100%)
(force: very likely 80-95%)

Figure 32. Automatic generation of a structured report from an argumentation.



What happened to the cesium-137 canister?

The cesium-137 canister likely (55-70%) was stolen. We assess it is very unlikely (5-20%) that the cesium-137 canister is being used in a project of the XYZ Company and that there is almost no chance (1-5%) that the cesium-137 canister was misplaced.

Damage to the locker where the cesium was stored and video surveillance of what appears to be the stolen canister being loaded into a truck at the XYZ warehouse the day before the cesium canister was discovered missing strongly suggest the canister was stolen.

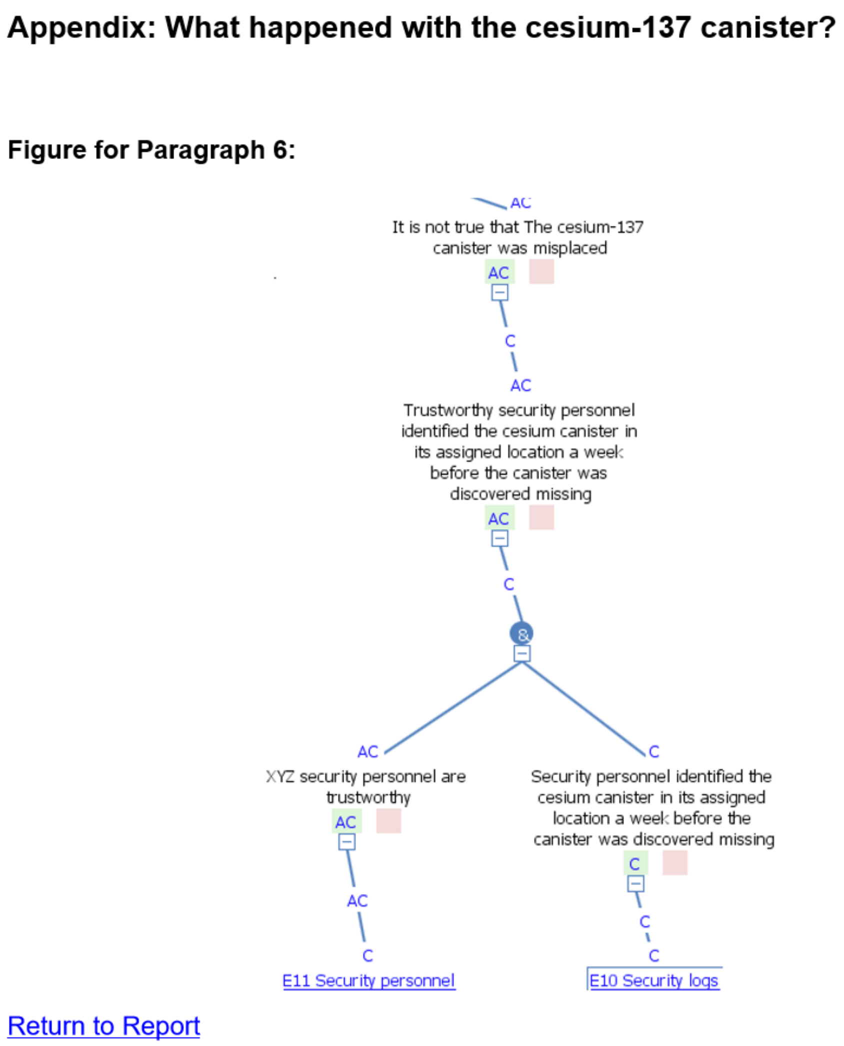
- After noticing that the lock containing the cesium canister appeared to have been tampered with, XYZ consulted a locksmith who confirmed that the lock was forced open.
- The security camera at the XYZ warehouse contains a video segment showing a person loading what is probably the stolen canister into a U-Haul panel truck one day before the canister was discovered missing. The record of the security guard indicates that a panel truck bearing Maryland license plate number MDC-578 was in the XYZ parking area that same day, suggesting that this was the vehicle used to steal the cesium.
Figure for Paragraph 3; Evidence for Paragraph 3

Grace, the Vice President for Operations at XYZ, reported that while they have several projects

Figure 33. Production report with links to corresponding argumentation fragments and evidence.

Figure 34 shows a fragment of the Appendix.

This process of creating reports simplifies the task of the user where they can focus on building and creating the argumentation and allow the system to write it as a semi-final product which already contains the majority of the information for their final analytic product. As opposed to the current state of the art (Google Docs), the user does not have to start with a blank slate. Time can be spent on argumentation development and refinement rather than on writing in order to improve the overall quality of the reasoning.



Appendix: What happened with the cesium-137 canister?

Figure for Paragraph 6:

It is not true that The cesium-137 canister was misplaced

Trustworthy security personnel identified the cesium canister in its assigned location a week before the canister was discovered missing

XYZ security personnel are trustworthy

Security personnel identified the cesium canister in its assigned location a week before the canister was discovered missing

E11 Security personnel

E10 Security logs

Return to Report

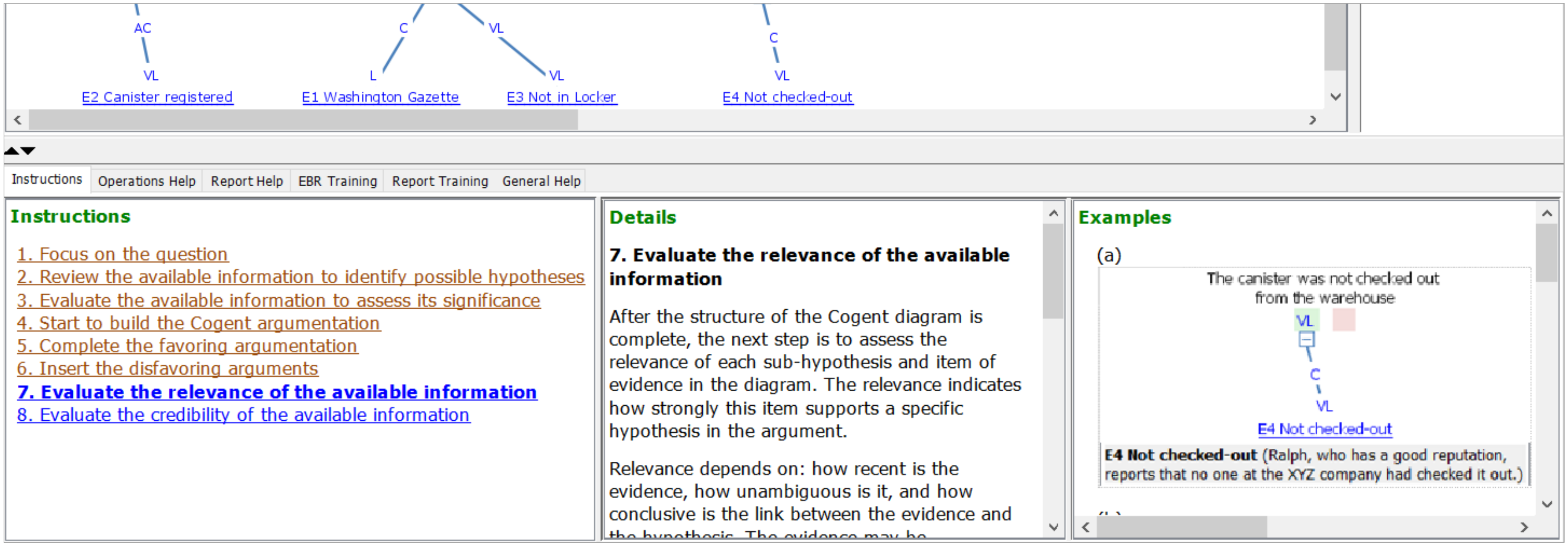
Figure 34. Appendix fragment showing an argument and the corresponding evidence.

3.3.6. Online Help

The entire Co-Arg training is also integrated into Cogent as online help. The help is accessed through the tabs situated at the bottom of the Whiteboard. Each help tab is briefly described below.

Instructions Tab

Under this tab there is a sequence of general instructions with guidance on how to develop an argumentation to answer an intelligence question. Figure 35 shows part of instruction "7. Evaluate the relevance of the available information." It briefly defines this relevance credential and exemplifies typical cases of relevance assessment.



Instructions

1. Focus on the question
2. Review the available information to identify possible hypotheses
3. Evaluate the available information to assess its significance
4. Start to build the Cogent argumentation
5. Complete the favoring argumentation
6. Insert the disfavoring arguments
7. Evaluate the relevance of the available information
8. Evaluate the credibility of the available information

Details

7. Evaluate the relevance of the available information

After the structure of the Cogent diagram is complete, the next step is to assess the relevance of each sub-hypothesis and item of evidence in the diagram. The relevance indicates how strongly this item supports a specific hypothesis in the argument.

Relevance depends on: how recent is the evidence, how unambiguous is it, and how conclusive is the link between the evidence and the hypothesis. The evidence may be

Examples

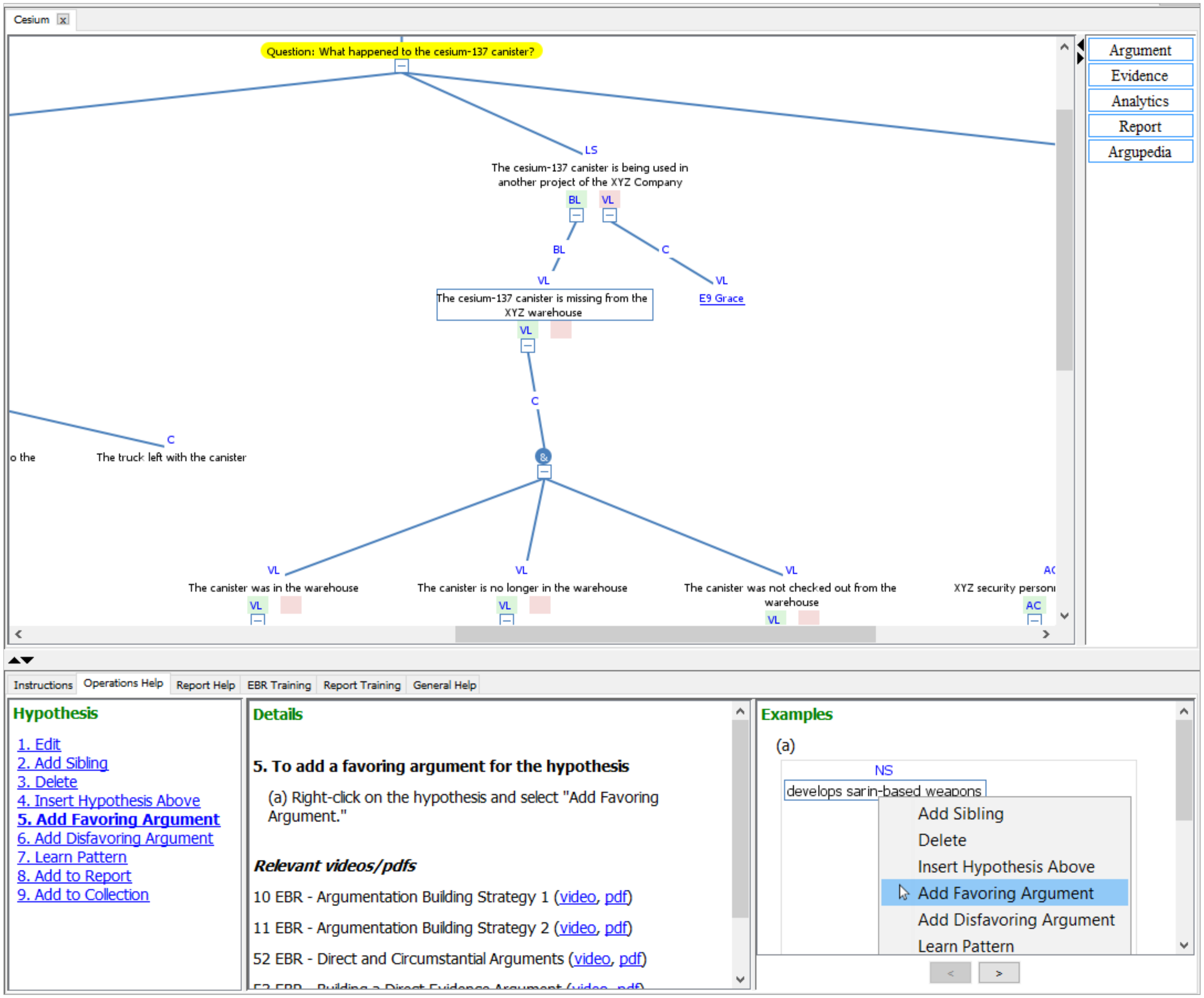
(a)

E4 Not checked-out (Ralph, who has a good reputation, reports that no one at the XYZ company had checked it out.)

Figure 35: General argument development guidance.

Operations Help Tab

This tab provides help on the operations that can be performed on the node selected in the argumentation. In the situation illustrated in Figure 36, the user clicked on a hypothesis. As a result, under the Operations Help tab the system displayed all the operations that can be performed on that node, in the bottom left pane. The user clicked on "Add Favoring Argument" and the system displayed the instructions on how to perform it in the bottom middle pane, also illustrating it in the bottom right pane. Notice that the middle pane also gives access to relevant videos and pdf documents from the training on evidence-based reasoning.


Cesium
Question: What happened to the cesium-137 canister?
LS
The cesium-137 canister is being used in another project of the XYZ Company
BL
VL
C
The cesium-137 canister is missing from the XYZ warehouse
E9 Grace
&
o the
The truck left with the canister
The canister was in the warehouse
The canister is no longer in the warehouse
The canister was not checked out from the warehouse
AC
XYZ security personi
Argument
Evidence
Analytics
Report
Argupedia
Instructions
Operations Help
Report Help
EBR Training
Report Training
General Help

Hypothesis

1. Edit
2. Add Sibling
3. Delete
4. Insert Hypothesis Above
5. Add Favoring Argument
6. Add Disfavoring Argument
7. Learn Pattern
8. Add to Report
9. Add to Collection

Details

5. To add a favoring argument for the hypothesis

(a) Right-click on the hypothesis and select "Add Favoring Argument."

Relevant videos/pdfs

10 EBR - Argumentation Building Strategy 1 (video, pdf)

11 EBR - Argumentation Building Strategy 2 (video, pdf)

52 EBR - Direct and Circumstantial Arguments (video, pdf)

53 EBR - Building a Direct Evidence Argument (video, pdf)

Examples

(a)

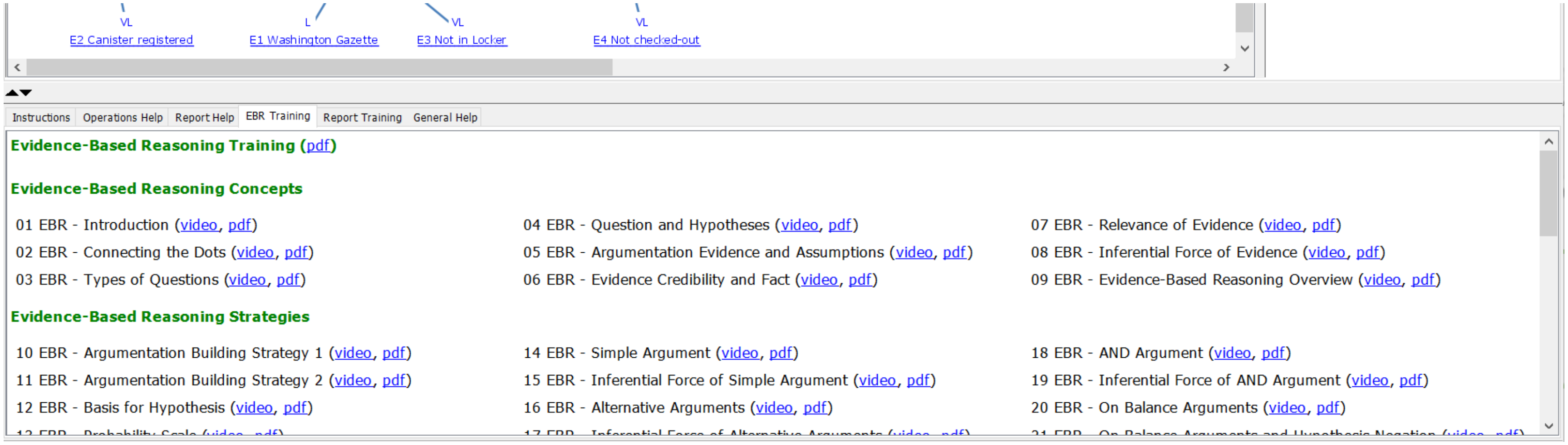
Figure 36: Context-dependent help.

Report Help Tab

The Report Help tab is similar to the Operations Help tab, providing context-sensitive help on the operations that can be performed on the report element selected in the Report Assistant.

EBR Training Tab

As illustrated in Figure 37, the EBR Training tab provides access to all the videos and pdf documents from the training on evidence-based reasoning.



Evidence-Based Reasoning Training (pdf)

Evidence-Based Reasoning Concepts

01 EBR - Introduction (video, pdf)
02 EBR - Connecting the Dots (video, pdf)
03 EBR - Types of Questions (video, pdf)
04 EBR - Question and Hypotheses (video, pdf)
05 EBR - Argumentation Evidence and Assumptions (video, pdf)
06 EBR - Evidence Credibility and Fact (video, pdf)
07 EBR - Relevance of Evidence (video, pdf)
08 EBR - Inferential Force of Evidence (video, pdf)
09 EBR - Evidence-Based Reasoning Overview (video, pdf)

Evidence-Based Reasoning Strategies

10 EBR - Argumentation Building Strategy 1 (video, pdf)
11 EBR - Argumentation Building Strategy 2 (video, pdf)
12 EBR - Basis for Hypothesis (video, pdf)
13 EBR - Probability Scale (video, pdf)
14 EBR - Simple Argument (video, pdf)
15 EBR - Inferential Force of Simple Argument (video, pdf)
16 EBR - Alternative Arguments (video, pdf)
17 EBR - Inferential Force of Alternative Arguments (video, pdf)
18 EBR - AND Argument (video, pdf)
19 EBR - Inferential Force of AND Argument (video, pdf)
20 EBR - On Balance Arguments (video, pdf)
21 EBR - On Balance Arguments and Hypothesis Negation (video, pdf)

Figure 37: EBR Training tab (videos and pdf).

Report Training Tab

Similar to the EBR Training tab, the Report Training tab provides access to all the videos and pdf documents from the training on report development.

General Help Tab

The General Help tab provides access to the general Co-Arg operations that are not included in the other tabs, such as "Submitting the Production Report".

4. User and Team Management

Argupedia allows basic operations for user and team management through the administrative portal, or implicitly at user login.

4.1. User Management

Argupedia keeps a minimal user profile consisting of:

1. User name (or nickname to be used in the system);
2. Email (if logging though the Argupedia portal);
3. Password-hash (if logging through the Argupedia portal);
4. User category (representing user's type and the experiment); For instance, *CBR CHOICE* is the category for the participating in the T&E experiment, and *2018 EBR Course GMU* is the category for the users participating in the internal experiments;
5. Assigned team(s).

To create users, Argupedia may employ the following methods:

1. Import from the T&E environment (if a user is already defined, its description is refreshed, if not, a new user is created in the Argupedia environment) – this feature was used in the T&E experiment;
2. Predefined in the system (for testing accounts and for administrative accounts);
3. Manually defined by an administrator – this feature was used in the internal experiments.

The following operations were allowed for user management in the administrative portal:

1. Visualization of the current users and their profiles (brief or detailed views);
2. Marking the training as done for a user (for testing purposes);
3. Adding a new user;
4. Changing the password for a user.

4.2. Team Management in the T&E Experiment

In the T&E experiments the teams were created dynamically, trying to maximize the number of teams while keeping a minimal team size.

The following prior observations were taken into account while deciding on the dynamic teaming method:

- *No Team Leader:* The team does not have a team leader. Any member of the team is allowed to independently continue the formal analysis and to submit a report. We included this option based on the feedback received from the initial internal experimentation, and the successful use of this teaming method in subsequent internal experiments.
- *Importance of First Session:* The members of a team do most of the work during their first session, and only review and have minor participation during the following sessions (based on anecdotal evidence). Therefore, the importance of when the first session takes place must be taken into account for team formation.
- *Tradeoff between Team Size and Number of Teams*: While we would have preferred to have larger teams to test the collaborative capabilities of Argupedia, we also needed to ensure a large enough number of teams. In the internal experiments we observed that very small teams may not be functional because of the lack of timely participation by some users.

Based on these observations we developed the following teaming method for the T&E experiment:

- For each problem we may have or not an open team (i.e., a team still accepting members for solving a given problem).
- When a user starts solving a problem, if the user is not already assigned to a team, s/he would be assigned to an open team (an existing team or a new one if none is open).
- A team is closed when one of the following conditions is satisfied:
    - A maximum number of users was reached (currently 12 users);
    - A minimum number of users was reached (currently 6 users), and a minimum teaming interval passed (currently 6 hours);
    - A required number of users was reached (currently 2 users), and a maximum teaming interval passed (currently 24 hours).

This method has the advantage to group the users that are starting to solve the same problem at approximately the same time, trying to create teams containing between 6 and 12 users. In order to avoid one-person teams, we introduced the requirement for a team to contain at least two users.

4.3. Team Management in the Internal Experiments and in the Pilot Experiments

In the internal experiments the teams were generated randomly at the beginning of the experiment and they were kept unchanged during the entire experiment. All the members in a team were required to continue to perform the formal analysis and to submit a report. The teams and their members' user accounts were created by an administrator using the team management tools.

In the pilot experiment, for the COARG condition, the teams were predefined at the beginning of the experiment and imported in Argupedia. In this condition we required one of the users to play the role of the lead analyst. In order to facilitate the selection of the leader we developed a leader negotiation component.

5. Training Requirements

5.1. Introduction

Using Co-Arg requires knowledge of the theory of evidence-based reasoning which is embedded into its functionality.

We have developed extensive training in evidence-based reasoning and the use of Co-Arg. However, based on the less optimal results of this training from the pilot study, and the feedback received from IARPA, we have significantly improved the training material, as well as the Argupedia module that controls the training, to make the process more adaptable to various user preferences and needs:

- Similar to the other performers, we developed a short video to encourage the volunteers to use our system. This video is accessible at: https://youtu.be/7_fuCELpUL0
- We divided the training into required and optional sections; the user had to complete the required part in order to access the problems in the experiment.
- We inserted multiple-choice questions with immediate feedback to reinforce the introduced concepts.
- We inserted hands-on training with Cogent much earlier in the training process, and we have organized it into short sessions.
- We provided shorter versions of the more demanding training parts (such as transforming the structured report into the production report), keeping the more complex versions as optional.
- We provided alternative ways of completing the training, either watching a video (with optional captioning) or reading an equivalent text.
- We incorporated all the training into the system as online help (see Section 3.3.6).

The total duration of the required training is about two hours. Optional training for an additional two hours is also available to help the users better understand and use the system and the theory underlining it.

The Argupedia Training Management module guides the user to watch all the required videos, or read the equivalent text. If a user attempts to work on a problem before completing the required training, s/he is directed to first finish the training from the point it was interrupted.

5.2. Required Training in Evidence-Based Reasoning

The topics of the required training are shown in Figure 38 and briefly described below.

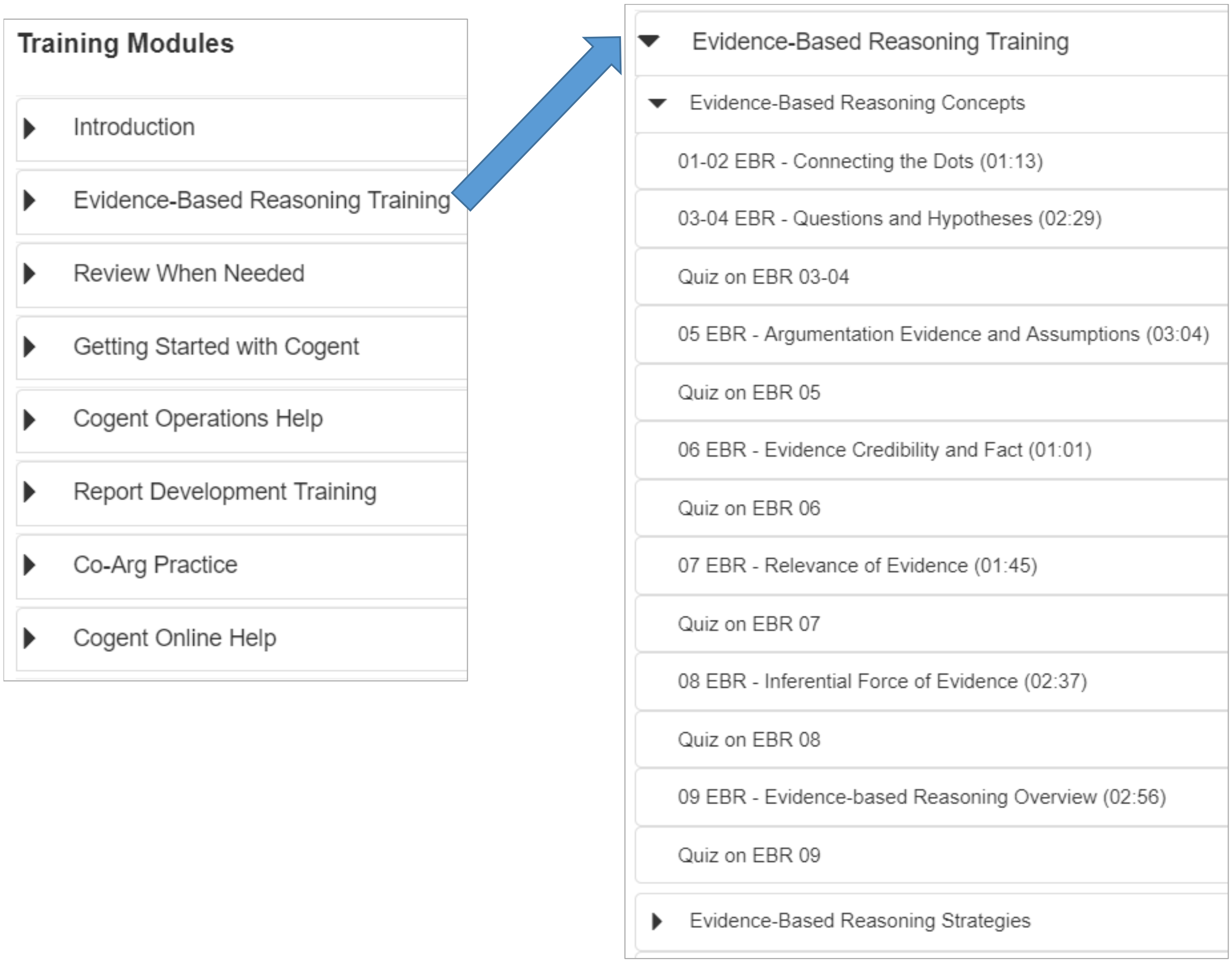

Training Modules
Introduction
Evidence-Based Reasoning Training
Review When Needed
Getting Started with Cogent
Cogent Operations Help
Report Development Training
Co-Arg Practice
Cogent Online Help
Evidence-Based Reasoning Training
Evidence-Based Reasoning Concepts
01-02 EBR - Connecting the Dots (01:13)
03-04 EBR - Questions and Hypotheses (02:29)
Quiz on EBR 03-04
05 EBR - Argumentation Evidence and Assumptions (03:04)
Quiz on EBR 05
06 EBR - Evidence Credibility and Fact (01:01)
Quiz on EBR 06
07 EBR - Relevance of Evidence (01:45)
Quiz on EBR 07
08 EBR - Inferential Force of Evidence (02:37)
Quiz on EBR 08
09 EBR - Evidence-based Reasoning Overview (02:56)
Quiz on EBR 09
Evidence-Based Reasoning Strategies

Figure 38: Required training.

Evidence-Based Reasoning Training: A sequence of short videos with the total duration of 36 minutes, interleaved with multiple-choice questions with immediate feedback on the presented material (see the right hand side of Figure 38). The user may opt to read a textual description instead of watching a video.

Review When Needed: Short videos (and equivalent pdf descriptions) explaining various operations with Co-Arg, such as “Submit Production Report.” The user does not need to review them at the time of training, but be aware of their existence.

Getting Started with Cogent: A step-by-step guided development of an argumentation with Cogent, intended to present the main available operations.

Cogent Operations Help: Brief introduction of the context-dependent help in Cogent (see Section 3.3.6).

Report Development Training: A 5-minutes video presenting how to generate a structured report in Cogent, how to insert argumentation fragments, and how to submit the report.

Co-Arg Practice: A sequence of three simple problems to be solved with Cogent, and their explained solutions.

Cogent Online Help: A 2-minutes video introduction of the Cogent online help (see Section 3.3.6).

5.3. Optional Training in Evidence-Based Reasoning

The topics of the optional training are shown in Figure 39 and briefly described below.

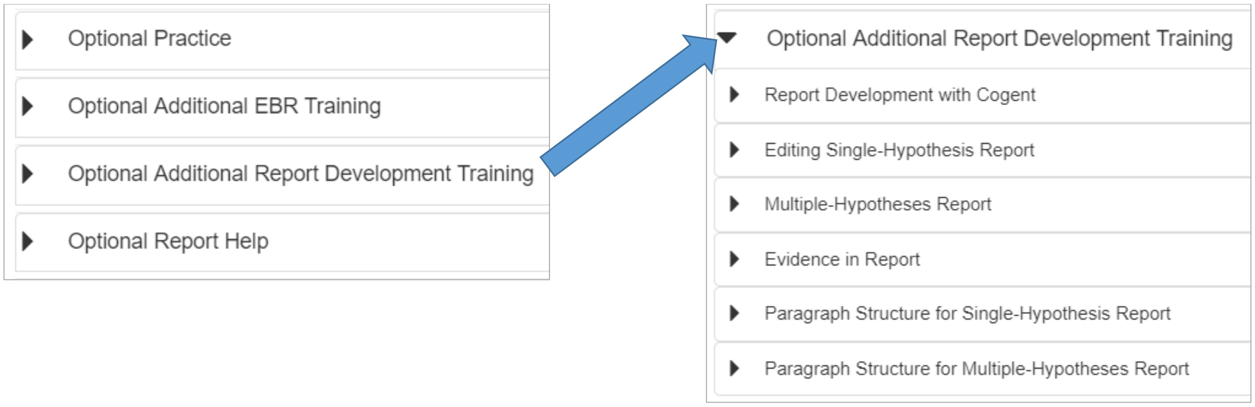

Optional Practice
Optional Additional EBR Training
Optional Additional Report Development Training
Optional Report Help
Optional Additional Report Development Training
Report Development with Cogent
Editing Single-Hypothesis Report
Multiple-Hypotheses Report
Evidence in Report
Paragraph Structure for Single-Hypothesis Report
Paragraph Structure for Multiple-Hypotheses Report

Figure 39: Optional training.

Optional Practice: A sequence of three complex problems to be solved with Cogent, and their explained solutions.

Optional Additional EBR Training: Additional training videos on evidence-based reasoning.

Optional Additional Report Development Training: A sequence of videos providing detailed training on how to transform a structured report into a more comprehensible and persuasive production report. The main parts of this training are shown in the right hand side of Figure 39.

Optional Report Help: Brief introduction of the context-dependent report help in Cogent

6. Problems Development

6.1. Problems Development Approach

Co-Arg is being developed to improve reasoning when addressing typical intelligence analysis problems. These problems require answering intelligence questions about a situation of interest by developing defensible and persuasive argumentations to assess the probability of alternative hypotheses based on the available incomplete, uncertain, ambiguous, contradictory, and missing information. The overwhelming majority of the problems encountered by the intelligence analysts in their practice are of this type.

We have developed evidence-based reasoning problems of the type intelligence analysts routinely encounter. These problems included:

- A description of a situation, an intelligence question to answer, and additional imperfect information to use in answering the question.
- Evidence supporting multiple hypotheses that needed to be marshaled and evaluated for both credibility and relevance to each hypothesis, the incorporation of logical inferences and assumptions to answer the question posed, and justification for assumptions with an assessment of their credibility.
- Information from different types of sources, including human sources, intercepted communications, documentary evidence, etc.
- Information obtained from human sources that included information that can be used to assess the source's reliability and access.
- Both favoring and disfavoring evidence for different hypotheses.
- The formulation of an answer to the intelligence question in the form of a production report.

The problems were developed to meet all 8, or at least the first 7 key elements of the high quality of analytic reasoning specified by IARPA and T&E. Additionally, we have developed different problems for different activities, with no overlap in problems, to avoid side effects, as follows:

- Problems submitted to be used in the T&E evaluation;
- Problems to be used in practice with Co-Arg (they include significant feedback);
- Problems to be used in internal testing and evaluation;
- Problems to be used in training on evidence-based reasoning and the use of Co-Arg.

6.2. Developed Problems

We have made a significant effort to develop many problems for use in our internal evaluations and in the T&E evaluations. We took special care not to use the same problem in different evaluations. We present these problems in the following.

The problems submitted to be used in the T&E evaluation are described in Table 3. Two of these problems have actually been used in the evaluation, "Fillistan Missile Test" (see Appendix 9.7) and "Who is the Spy?" (see Appendix 9.9).

Table 4 presents the problems used in the Fall 2017 evaluation with students in a class at CSUSB. Notice that none of these problems were used in the follow-on evaluations. "Economics Minister Resigns" and "What Did Jackson Decide?" were used as practice problem in the Spring-summer 2018 internal evaluation.

Problem Name	Question	Comments
Fillistan Missile Test	What kind of missile was launched from Fillistan on 10 January 2017?	This is one of the two most complex problems. It meets all the 8 key elements of high-quality analytic reasoning, including "generation of unique analytic insights." It was used in the T&E evaluation.
Who's the Spy?	Is Tom, Dick, Harry, or Paul the person removing the classified information from C/M-2 that has been passed to Razmania?	This is other one of the two most complex problems. It meets all the 8 key elements of high-quality analytic reasoning, including "generation of unique analytic insights." It was used in the T&E evaluation.
New Allegations of Forced Labor	Is forced labor again being used at the Goldplus gold mine?	Meets the first 7 of the 8 key elements of high-quality analytic reasoning, especially "assessment of probability" where "care should be taken to avoid traps of over-specification and over-confidence."
Crossing the Border	Did Pomania cross the LOO into Tussia-controlled territory in Renmark and initiate the fighting with Tussia on 5 July 2017 as claimed by Tussia?	Meets all the 8 key elements. One element of the analysis requires creativity and imagination to interpret some information that is diagnostic but which could easily be ignored as non-diagnostic. It was used in the T&E evaluation.
Is Martin Corrupt?	Did President Martin approve the BCC contracts to enrich his son Mike Martin despite knowing that the cost of the contracts were grossly inflated and unjustified?	Meets 7 of the 8 key elements of high-quality analytic reasoning, especially "clear marking and justification of key judgments" and the "assessment of quality, credibility, and diagnosticity of evidence." It was used in the T&E evaluation.

Table 3: Problems submitted to be used in the T&E evaluation.

Problem Name	Question
The Complex Delta Mystery	What function does Complex Delta have?
Economics Minister Resigns	Why Economics Minister Handle had to resign?
Informant Evaluation	Why is Tamar Zulat offering this information?
Questionable Activity at Platinum Mine	Is forced labor again being used at the Agadir platinum mine?
What Did Jackson Decide?	Will President Jackson decide to change Cartia's nuclear-launch protocol for its land-based missiles to launch on warning, i.e. as soon as Cartia's satellites and other sensors detect a launch?

Table 4: Testing problems used in the internal evaluation in Fall 2017.

Table 5 presents the testing problems used in the internal evaluation in Spring-Summer 2018. Notice that for each problem we defined a clone problem that requires the same logic but looks different.

Problem 1: Fighting Erupts in Midland	**Clone 1:** Adversarial Activities in Victorland
Question: Did Wokistan cross the LOS into Transvindia-controlled territory in Midland and initiate the fighting with Transvindia on 5 July 2017 as claimed by Transvindia?	**Question:** Did Upickastan cross the LOS into Maranathia-controlled territory in Victorland and initiate the fighting with Maranathia on 5 July 2017 as claimed by Maranathia?
Comment: Both problems are clones of “Crossing the Border” proposed for T&E evaluation.	
Problem 2: Mortar Sale	**Clone 2:** Chemical Precursor Deal
Question: Did President Jones of Stanistan approve this mortar sale to Badlandia?	**Question:** Did President Thomas of Tuslia approve the sale of chemical precursors to Postonia?
Problem 4: Bomber Crashes at Take Off	**Clone 4:** Tank Explodes During Testing
Question: What kind of bomber was flight tested in Wakanda on 10 January 2017?	**Question:** What tank was conducting live-fire exercises at Basrana’s Tank Testing Grounds on 10 January 2017?
Problem 5: Euclid Makes Offer to Bokota	**Clone 5:** Brineland Receives Unexpected Information
Question: Is Bill, Chris, Joe, or Mike the person removing the classified information that has been passed to Bokota from D/BT?	**Question:** Is Sean, Zack, Hugh, or Joshua the person removing the classified information that has been passed to Brineland from DCR/5?
Comment: Both problems are clones of “Who’s the Spy?” proposed for T&E evaluation.	

Table 5: Testing problems used in the internal evaluation in Spring-Summer 2018.

The problems used for training on evidence-based reasoning and the use of Co-Arg are presented in Table 6.

Problem Name	**Question**
Hakka	Does Hakka have chemical weapons?
Cesium-137	What happened to the cesium-137 canister?
Stolen car	Did John steal the car?
Economics Minister Fired	Why did the Prime Minister fire the Economics Minister?
Entrance Exam	How will Nick do on the law school entrance exam?

Table 6: Problems used in training on evidence-based reasoning and the use of Co-Arg.

Finally, the problems used for practice with Co-Arg are presented in Table 7.

Problem Name	Question	Comments
Salazar	Is John Ventura, the director of Polombia's intelligence service, involved in the drug sale to Markistan that is being arranged by Joe Salazar?	Used as practice problem in both internal evaluations and in the T&E evaluation
Manada SAM Sale	Which SAM system is Manada selling Sindia?	Used as practice problem in both internal evaluations and in the T&E evaluation
Economics Minister Resigns	Why did Handle resign or was asked to resign?	Used as testing problem in the Fall 2017 internal evaluation
Real Estate Scam	Did President Sanchez approve the purchase of the GIR properties as part of a corrupt scheme to enrich his son Juan Sanchez?	
Jackson Decision	Do you agree with the position that President Jackson will not change Cartia's nuclear-launch protocol for its land-based missiles to launch on warning?	Used as testing problem in the Fall 2017 internal evaluation

Table 7: Problems used for practice with Co-Arg.

7. Internal Experimentation and Evaluation

7.1. Design and Methods

The internal research experiment described in this section was pre-registered with the Center for Open Science (https://osf.io/r2f8s/, Appendix 9.1). We evaluated the impact of using Co-Arg as well as its associated training by recruiting students enrolled in four courses across four universities: California State University at San Bernardino (CSUSB), University of Mary Washington (UMW), George Mason University (GMU), and University of Nebraska at Omaha (UNO). Those students who volunteered to participate used two different systems, Google Docs and Co-Arg, to aid in producing written responses to intelligence questions. In addition, students received training in evidence-based reasoning and using each respective system. This experiment was used to evaluate different aspects of Co-Arg and its associated training. Each aspect is summarized in a different section. First we describe the methods which were common to all studies.

7.1.1. Experimental Design

The purpose of the study was to evaluate the value of Co-Arg tools and training during the process of producing written reports to intelligence problems. We evaluated the performance of individuals when using Co-Arg with team support. As a control condition we compared individual performance using Co-Arg to performance using Google Docs, which provides only general-purpose collaborative tools and no reasoning tools. In contrast, Co-Arg provides collaborative tools specialized for intelligence problems and reasoning support.

Participants were assigned to teams of 3 to 6 members at their University and remained in the same teams for the duration of the school term. For each problem participants individually produced a written report in response to the intelligence question. However, regardless of the system, for each problem there was an initial period of several days in which they shared ideas as a team. Thus, individuals' performance may be in part due to team-level factors.

This study used a repeated measures, within-subjects design. Each participant was tested under both of the two main conditions. Table 8 summarizes the design. The control condition took place before participants were trained using Co-Arg. In this condition participants used the control system Google Docs. The experimental condition took place after participants were trained using Co-Arg. In this condition participants used the experimental system Co-Arg. For each condition participants worked on two intelligence problems and produced a written report for each problem. The design was further nested because individuals belonged to and shared ideas within their teams.

	Condition (within-subjects)	
	Control/Pre-test (before training, using control system Google Docs)	**Experimental/Post-test** (after training, using experimental system Co-Arg)
Number of observations per participant	2 problems	2 problems

Table 8: Summary of the experimental design.

We chose not to counterbalance the order of the two conditions. That is, all participants were tested under the control condition before being tested under the experimental condition. We chose not to counterbalance the order because we believe that the training provided to learn how to use Co-Arg helps to improve reasoning about intelligence problems regardless of whether Co-Arg is used for a given problem. Thus, testing under the control condition had to happen before training was applied.

Problems were counterbalanced so that all 4 problems were used in both conditions (control, experimental) and each participant worked on a problem only once. 2 of the problems were single-hypothesis problems while the other 2 were multiple-hypotheses problems. In each condition there was 1 single-hypothesis problem and 1 multiple-hypotheses problem. Each team worked on the same problem at the same time. No solutions were given out before completion of the experiment and participants were asked not to discuss the problems outside of their teams. In addition, for each problem we created a clone, which only differed superficially (e.g., different names for countries, individuals). See Table 9 for problem allocation.

University	Team	Google Docs		Co-Arg	
		Pre-test 1	Pre-test2	Post-test 1	Post-test 2
UMW	1	5	1	2	4
CSUSB	1	2	4c	1	5c
	2	4	2c	5	1c
	3	1	5c	2	4c
	4	5	1c	4	2c
UNO	1	2	4c	1	5c
	2	4	2c	5	1c
	3	1	5c	2	4c
	4	5	1c	4	2c
	5	5c	1	2c	4
GMU	1	2	4c	1	5c
	2	4	2c	5	1c
	3	1	5c	2	4c

Table 9: Counterbalanced problems across the teams. 1 = Fighting Erupts in Midland, 2 = Mortar Sale, 4 = Bomber Crashes at Take Off, 5 = Euclid Makes Offer to Bogota. 1 and 2 are single hypothesis problems and 4 and 5 are multiple hypothesis problems. Postfix c indicates a clone problem. Problem 3 was initially intended to be used for testing but it was no longer needed and was used for practice.

7.1.2. Participants

The population sample is characterized by young adults, English-speakers, following post-secondary education in U.S. institutions. The participants were recruited through purposeful sampling (inclusion criterion: following a specific course).

Participants were recruited from four universities: California State University at San Bernardino (CSUSB), University of Mary Washington (UMW), George Mason University (GMU), and University of Nebraska at Omaha (UNO). Students in the graduate level National Security Studies program at CSUSB, students in the undergraduate anthropology major at UMW, graduate students in Information Sciences and Technology, as well as undergraduate students in Criminology, Law and Society at GMU, and undergraduate and graduate students in Political Science and International Studies, Emergency Management, Interdisciplinary Informatics and Cybersecurity at UNO, were invited to participate. Students were offered the opportunity to participate in the study in exchange for an independent-study or a special topics course credit and the knowledge and experience gained from the course. Following IRB guidelines, participation was not required and students were provided with alternative assignments if they chose to no longer participate in the study.

Students CSUSB were enrolled in SSCI 695-01 for 4-credits, a quarter graduate level Independent Studies course in the National Security Studies or National Cyber Security Studies program. Students at UMW were enrolled in URES 197 for 2-credits, a semester undergraduate level independent participatory research class, or they were volunteers. Students at GMU were enrolled in AIT 499/CRIM 490 for 3 credits, a summer special topics course. Students at UNO were enrolled in PSCI 4920/8926 for 3 credits, another summer special topics course.

Independent Study Course 1

Graduate students enrolled in the National Security Studies Program California State University at San Bernardino (CSUSB). Students were offered the opportunity to earn 4-credits for participation in the study if they enrolled in SSCI 695-01. Twenty CSUSB students initially enrolled in the course, 18 of whom were enrolled in the Master of Arts in National Security Studies, and 2 of whom are enrolled in the Master of Science in National Cyber Security Studies. Of these 20 students, 8 are females and 12 are males. During week 1 of the course, 1 student dropped from the course.

Independent Study Course 2

Undergraduate students at University of Mary Washington (UMW). Students were offered the opportunity to earn 2-credits for participation in the study if they enroll in URES 197. However, of the four students two were volunteers. The students who have elected to participate at UMW include only female undergraduate anthropology majors (both male and female students were recruited, but most anthropology majors at UMW are female).

Special Topics Course 3

Undergraduate and graduate students enrolled in AIT 499/CRIM 490 at George Mason University. This is a 3-credit course cross-listed between the Department of Information Sciences and Technology of the Volgenau School of Engineering (AIT 499) and the Department of Criminology, Law and Society of the College of Humanities and Social Sciences (CRIM 490).

Special Topics Course 4

Undergraduate and graduate students enrolled in PSCI 4920/8926: Special Topics National Security and Intelligence Practicum at the University of Nebraska at Omaha. Students were given the opportunity to earn 3-credits for participation. The students came from the College of Arts and Sciences (Department of Political Science and International Studies major), College of Public Affairs and Community Service (Department of Emergency Management), and College of Information Science and Technology (School of Interdisciplinary Informatics and Cybersecurity major).

7.1.3. Participant Demographics

Sixty two students were recruited and consented to participate in the study (UMW: 4, CSUSB: 19, UNO: 28, GMU: 11). These participants were assigned to a total of 13 teams between 3 and 6 members (Mean = 4.8). 65% of participants were male.

62% of participants were enrolled in a graduate program, while 38% were enrolled in an undergraduate program. The students were studying political science (26%), national security studies (21%), Cybersecurity (21%), IT or Information Systems (15%), Criminology (11%), the rest had a variety of majors (e.g. International Studies, English).

Six students (9.7%) had prior experience working with an early version of Co-Arg in a Fall 2017 class. 37% reported some form of experience with evidence-based reasoning (e.g., debate class, on the job experience as police officer).

7.1.4. Inclusion and Exclusion Criteria

Participants were included if they met two criteria: 1) they completed at least one test problem using both systems and 2) they at least partially completed two practice problems using Co-Arg. Participants could not be included in statistical analysis if they failed to use both systems as this would not make it possible to compare quality of reasoning or quality of communication given the within subjects design. Participants were included in analyses even if they failed to submit a report for 1-2 of the test problems, however observations for each test problem that they failed to complete were omitted (rather than marked as 0). Without enough practice participants would not be able to effectively use Co-Arg. Participants were assigned to complete 5 practice problems using Co-Arg; we used a lax inclusion criterion of only requiring 2 practice problems.

A majority of participants, 76%, submitted reports for every test problem. Using Google Docs 94% submitted reports for the first problem and 95% for the second problem. Using Co-Arg 81% submitted reports for the first problem and 85% submitted reports for the second problem. Fewer reports were submitted for Co-Arg because at the time of testing more participants had dropped out of the study permanently.

52% of participants (at least partially) completed 5 practice problems using Co-Arg, 26% completed or partially completed only 4 practice problems, 5% completed or partially completed only 3 practice problems, 3% completed or partially completed only 2 practice problems, 5% completed or partially completed only 1 practice problem, and 10% did not even partially complete any practice problems.

9 participants (15%) failed to complete at least one test problem with both systems. Typically, these students dropped out of the study prior to using Co-Arg on test problems. This included all 4 students from UMW. 9 participants (15%) failed to partially complete at least two practice problems. When both inclusion criteria were applied this left 51 students (11 students were excluded, 18%). Only one team, the team from UMW, had to be excluded because all of its members were excluded. The other 12 teams were included, these teams had between 66% and 100% of their original members and had between 3 and 6 members.

Most results summarize only those data for the 51 participants in 12 teams that were ultimately included in the study.

7.1.5. Procedure

The experiment took place over around 10 weeks per course.

Pre-test Training (1 day)

The pre-test training introduced the types of problems that would be worked on, the expected solutions (production reports) and how they would be evaluated, as well as how to use Google Docs for asynchronous collaboration. Informed consent was provided and a demographic survey was completed.

Pre-test (2 weeks, one problem/week)

The participants, organized in teams, collaborate asynchronously, until noon of Day 3, to brainstorm and develop an initial solution, by working on a shared Google Doc that is provided to them. When this deadline is reached, they will no longer have rights to modify the jointly developed document, but may continue to view it. At the same time, each will receive a personal Google Doc that will be a copy of what they wrote together. They will work independently, each finalizing the solution, until Day 7 of the week. This mirrors the use of Co-Arg during the post-test, where at the beginning each team uses Argupedia for asynchronous collaboration, and then users independently finalize their solutions using Cogent.

This concludes the control/pre-test condition, where the participants solve problems using only Google Docs.

Training (1 week)

The participants are provided training in evidence-based reasoning and report development with Co-Arg, and in the use of the two components of the system, Argupedia and Cogent. The training is in the form of videos that they watch by themselves, as well as videos that direct them how to use the system.

Practice Problems (5 weeks, one problem/week)

At the beginning of each week, each team receives a practice problem. They are asked to use Argupedia in each of the first two days to contribute to the development of the informal analysis and review the contributions of the other members of the team. At the beginning of Day 3 each user reviews the informal analysis developed by the team and imports the desired results into Cogent. Each user, individually, continues the development of the solution in Cogent, develops the production report, and submits it by the end of Day 6. During Day 7 they receive the solution, an explanation of the solution, and an evaluation grid, and are asked to study them. During a group meeting, they discuss the solution with the instructor.

Post-test (2 weeks, one problem/week)

Each week each team is given a post-test problem to solve with Co-Arg. The process is similar to that from the pre-test, except that they use Co-Arg to solve the problems. During each of the first two days they use Argupedia to contribute to the development of the informal analysis and review the contributions of the other members of the team. At the beginning of Day 3 each user reviews the informal analysis developed as a team and imports the desired results into Cogent. Each user, individually, continues the development of the solution in Cogent, develops the production report, and submits it by the end of Day 7.

This concludes the treatment/post-test condition.

Final Discussion (1 day)

A group meeting during which the users fill-in several questionnaires and discuss the experiment with the instructor.

7.1.6. Materials

Participants completed the four questionnaires described below.

A demographic questionnaire completed once prior to the study which asked socio-demographic characteristics, such as degree program, major, and gender.

A post-problem questionnaire completed nine times, once after each test problem and each practice problem, which asked about the time spent on working on the problem broken down by phase (e.g., as a team, as an individual), a self-reported evaluation of the usefulness of the ideation tool (Google Docs or Argupedia), and open-ended question asking about any issues experienced using the tools.

A training questionnaire completed once after the participant had finished the training, which asked about time spent on various parts of the training, whether participant understood key concepts of evidence-based reasoning and whether the participant knew how to use basic functions of Co-Arg.

A final questionnaire completed once after all problems had been finished, it asked participants to self-report the helpfulness or unhelpfulness of each system at improving quality of reasoning and quality of communication, as well as measures to evaluate the usability of Co-Arg and of its components (Argupedia, Cogent) using the System Usability Scale (See Appendix 9.2) and the Net Promoter Score (See Appendix 9.3).

Table 10 shows some examples of items from these questionnaires.

Example Questionnaire Items
Thinking back on your experience working on the problem this week. Did you experience any issues with your tools or collaboration that made it more difficult to solve the problem? Please explain.
I was able to share my ideas and to be understood during the brainstorming session this week.
I was able to understand others' ideas during the brainstorming session this week
I used some of my teammates' ideas in developing my own solution to this week's problem
I think that [Co-Arg, Google Docs] improved my quality of reasoning when solving an intelligence problem (for example, by helping me think about and solve an intelligence problem).
I think that [Co-Arg, Google Docs] improved my quality of communication when writing a report describing the solution of an intelligence problem (for example, by helping me record my solution with more clarity and completeness that others could more easily understand it).
How has your reasoning changed or not changed as a result of working on these intelligence questions with Co-Arg? If your reasoning has changed please explain how it has changed and include concrete examples.

Table 10: Example Questionnaire Items.

System Usability Scale (SUS)

The System Usability Scale is a 10-item questionnaire designed to evaluate the usability of a system. It employs a 5-point Likert scale with semantic differential ranging from strongly agree to strongly disagree. The items assessed are:

- I think that I would like to use this system frequently.
- I found the system unnecessarily complex.
- I thought the system was easy to use.
- I think that I would need the support of a technical person to be able to use this system.
- I found the various functions in this system were well integrated.
- I thought there was too much inconsistency in this system.

- I would imagine that most people would learn to use this system very quickly.
- I found the system very cumbersome to use.
- I felt very confident using the system.
- I needed to learn a lot of things before I could get going with this system.

Net Promoter Score (NPS)

The Net Promoter Scores measures the probability of a system to be recommended for use in a scale ranging from -100 to +100. It assesses the satisfaction with a product or the loyalty of the users towards a system. The results are given by a percentage value. The survey rates whether the users are promoters, passive or detractors. While promoters tend to be satisfied with the solution, detractors are less likely to recommend it to other users. Passive users have a neutral response to the product.

The participants answer the following question:

"On a scale of 0-10, how likely is it that you would recommend Co-Arg to other people?"

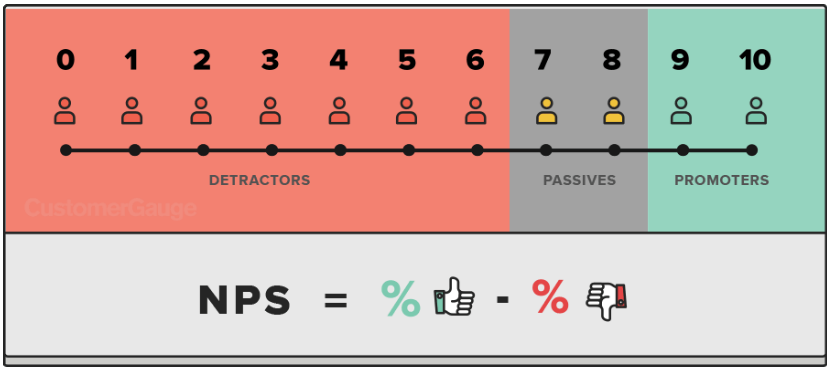


Figure 40: Classification of the respondents' answers with the Net Promoter Score.
Source: https://customergauge.com/blog/how-to-calculate-the-net-promoter-score/

7.2. Study: Assessment of Co-Arg in Improving the Quality of Evidence-Based Reasoning

We found that using Co-Arg and its associated training improved quality of reasoning scores on intelligence problems by 0.77 points out of 10. This improvement in quality of reasoning corresponds with a medium effect size of 0.48. A more detailed examination shows that the gains are due to improvement in the evaluation of sources of evidence and improvement in the argumentation structure.

7.2.1. Study Goal and Key Research Questions

This study served as a system evaluation. We evaluated whether Co-Arg and its associated training in evidence-based reasoning improved quality of reasoning in response to intelligence problems. In this study we investigated one primary research question to evaluate the effect of using Co-Arg on quality of reasoning; this research question and hypothesis were pre-registered with Center for Open Science (Original Registration: https://osf.io/r2f8s/, Amendment: https://osf.io/e3bdu/).

Research Question 1: *Does Co-Arg with its associated training improve quality of reasoning on intelligence problems for all users?*

Hypothesis 1: *We hypothesized that on average all participants would generate written reports that would score higher on quality of reasoning in response to post-test problems using Co-Arg compared to pre-test problems using Google Docs.*

In our analysis of Research Question 1 we deviated slightly from the pre-registered analysis plan in how we excluded participants. We did not anticipate participants dropping out before using both systems, however several participants did drop out before using Co-Arg. Although mixed-effect regression allows for the inclusion of repeated measures with some missing values, we decided to exclude participants who did not submit at least one test problem with both systems (Google Docs and Co-Arg). We made this decision because individual differences had a large effect on quality of reasoning scores and our goal was to compare individuals' scores using both systems. This decision meant we excluded an additional 4 reports; we conducted sensitivity analysis to show that this decision did not affect our results.

We followed up on this primary research question with four additional research questions to deepen our understanding of how Co-Arg improved quality of reasoning, whether students perceived an increase in quality of reasoning, for whom Co-Arg improved quality of reasoning, and for which problems Co-Arg improved quality of reasoning. These follow-up research questions 2-5 were exploratory and not pre-registered.

Research Question 2: *Along which dimensions of quality of reasoning, if any, does Co-Arg improve quality of reasoning on intelligence problems?*

1) *Hypothesis generation and accuracy of solution*
2) *Argumentation structure and reasoning*
3) *Identification of sources and assessment of credibility of evidence*
4) *Identification of key missing information and assumptions*

Research Question 3: *Do students perceive that their quality of reasoning has improved when using Co-Arg? If so, how?*

Research Question 4: *Does Co-Arg with its associated training improve quality of reasoning for some users more than others?*

a. *Does it improve reasoning more for students who received more feedback during practice from their instructor?*
b. *What characteristics are in common for those participants who scored highly on quality of reasoning using Co-Arg and showed the most improvement?*

Research Question 5: *Does Co-Arg with its associated training improve quality of reasoning for some problems more than others?*

7.2.2. Methods

Data for this study came from the main experiment. We explain the study design, participants, procedures, materials in more detail in Section 7.1. 62 individuals participated in this study. Of these participants 51 individuals in 12 teams were included in the final analysis (see Section 7.1). To be included participants had to 1) submit at least one report for a test problem using both systems and 2) at least partially complete two practice problems. In addition, while data was included for all individuals meeting these criteria, quality of reasoning scores were only included for reports that were submitted. These scores were omitted because in most cases the student had not attempted to solve the problem and thus do not reflect the degree to which the system helped (or did not help) a student with reasoning.

Participants were assigned two pre-test problems using Google Docs and two post-test problems using Co-Arg. In response to these problems 199 reports were submitted, 100 pre-test reports and 99 post-test reports.

Each written report was scored by two independent raters using the quality of reasoning rubric developed for each problem (see Appendix 9.6 for an example of a grid). Each Quality of Reasoning Rubric consisted of four characteristics or dimensions of well-reasoned reports:

1) Hypotheses generation and accuracy of solution (Hypotheses);
2) Argumentation structure and reasoning (Argumentation);
3) Identification of sources and assessment of credibility of evidence (Sources);
4) Identification of key missing information and assumptions (Assumptions).

Raters evaluated each quality of reasoning dimension separately. A total quality of reasoning score was calculated for each rater by summing the subscores for each dimension.

The two independent raters reached high inter-rater reliability on the composite quality of reasoning score ($ICC_{overall}$ = 0.85, 95% CI: 0.80-0.88).

On the first three dimensions raters achieved high inter-rater reliability:

- Hypothesis generation and accuracy of solution ($ICC_{hypotheses}$ = 0.85 95% CI: 0.80-0.88);
- Argumentation structure and reasoning ($ICC_{argumentation}$ = 0.82, 95% CI: 0.77-0.87);
- Identification of sources and assessment of credibility of evidence ($ICC_{sources}$ = 0.82, 95% CI: 0.77-0.86)

On the fourth dimension the raters achieved moderate inter-rater reliability:

- Identification of key missing information and assumptions ($ICC_{assumptions}$ = 0.63, 95% CI: 0.52-0.72).

When the quality of reasoning subscores for any dimension differed by more than 20% a third rater scored the report. 152 subscores out of 396 had to be adjudicated, that is 38.4% of all subscores. A single subscore for a particular dimension and a particular report was calculated by taking the average of the two raters if no adjudication was needed; by taking the average of the two closest scores if an adjudicator was needed; or taking an average of all three scores if the adjudicator's score fell within 20% of the other two scores. A single total quality of reasoning score was calculated for each report by summing the four subscores for that report. Final scores were linearly transformed to a 0-10 point scale to normalize across problem rubrics with different scales (e.g., Mortar problem had a rubric with 30.5 possible points while Midland problem had a rubric with 40 possible points).

Statistical analyses were conducted in R using lme4 and lmerTest packages.

7.2.3. Results

Research Question 1: *Does Co-Arg with its associated training improve quality of reasoning on intelligence problems for all users?*

We conducted a mixed effects linear regression to evaluate the effect of being trained and using Co-Arg on quality of reasoning. We used the same statistical model as was proposed in the pre-registration. Quality of reasoning scores for reports generated by students were treated as the dependent variable. Which system, Co-Arg or Google Docs, they used to work on the problem was treated as the independent variable of interest (System). Given the repeated measure design we included random effects in the model. Team was included as a random effect because for all problems students brainstormed as part of a team prior to working on the problem independently. Problem was included as a random effect because we counterbalanced the order in which problems were worked on. User was included as a random effect because each student worked on problems using both systems. The data met all assumptions of the statistical test, with the exception of normality. Deviation in normality of residuals was minimal and mixed effects linear regression is robust against minimal deviations from normality.

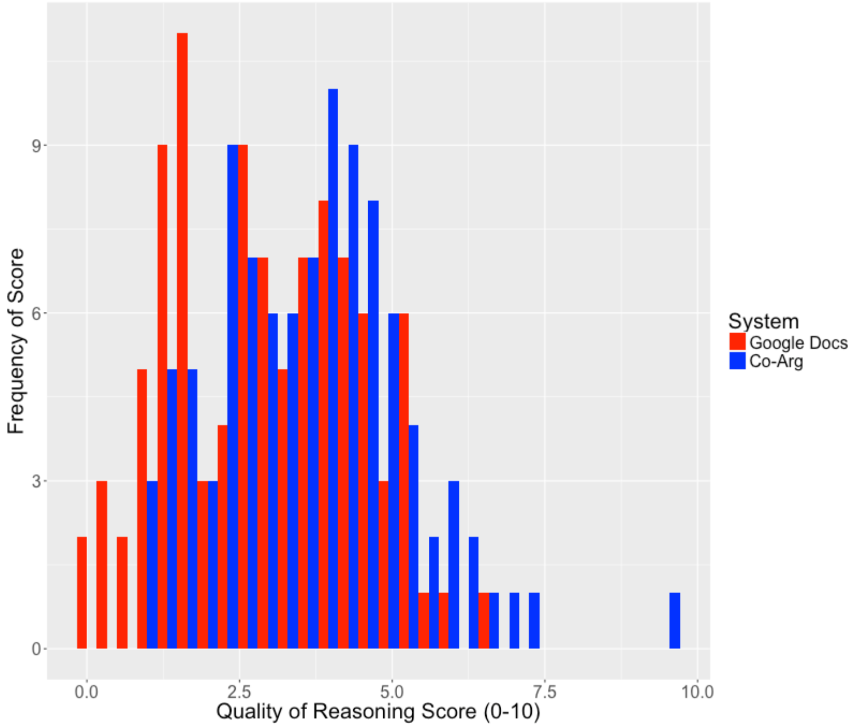


Figure 41: Histogram of quality of reasoning scores for problems solved with Google Docs and Co-Arg.

We found a significant effect of system usage on quality of reasoning scores. On average students scored 0.77 points higher on the 10 point measure of quality of reasoning when using Co-Arg compared to when using Google Docs. On average students scored 2.89 on quality of reasoning when solving intelligence problems with Google Docs and 3.66 when solving intelligence problems with Co-Arg. *This improvement represents a 0.48 standardized effect size (95% CI: 0.23 to 0.74)*.

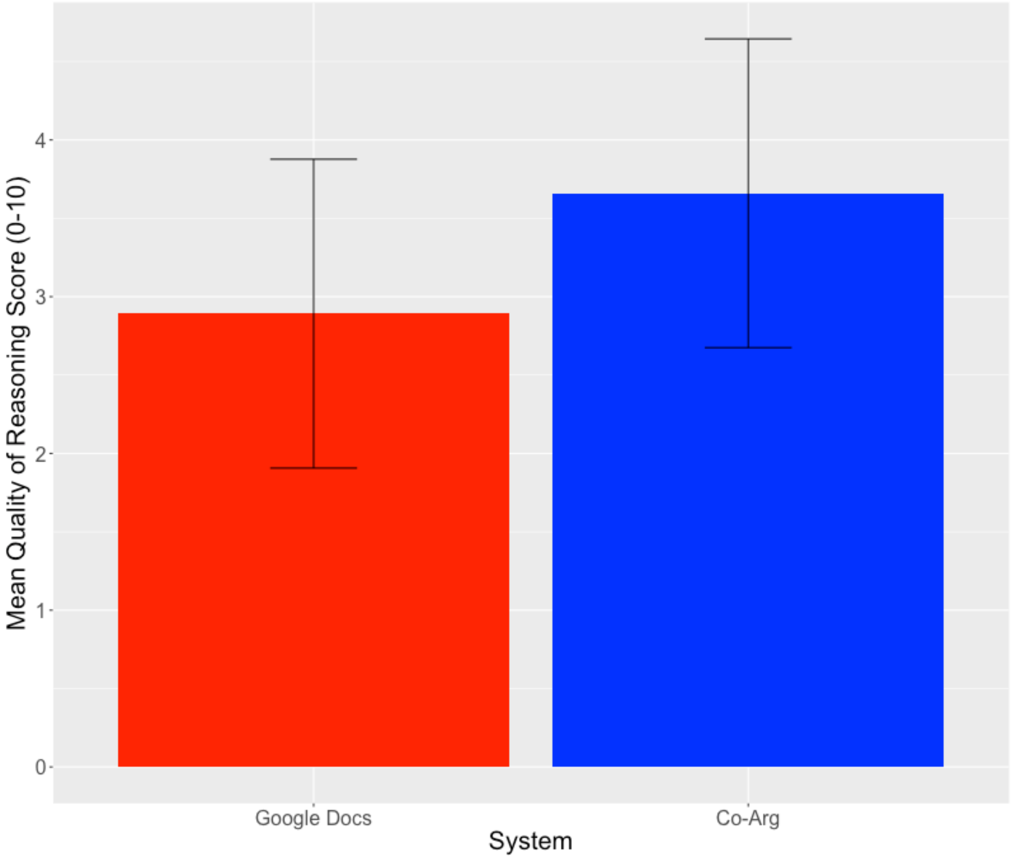


Figure 42: Estimated marginal means and 95% confidence intervals for quality of reasoning scores per system for the main mixed effects linear regression model.

	Coef.	SE	t	df	p	Effect Size
Intercept	2.89	0.37	7.86	4.390	0.0009	
System (Co-Arg vs. Google Docs)	0.77	0.19	3.97	146.03	0.0001	0.48

Table 11: Results of mixed effects linear regression model evaluating the effect of the system on quality of reasoning with user, team, and problem as random effects. System was treated as a factor with Google Docs as the base level. 199 scored reports from 4 problems generated by 51 users in 12 teams. Variance for random effects were user = 0.12, team = 0.15, problem = 0.41, error = 1.86.

System	Problem Order	Marginal Mean (95% CI)	Marginal Mean (95% CI)
Google Docs	1	2.82 (1.85-3.79)	2.89 (1.91-3.88)
	2	2.96 (1.99-3.93)	
Co-Arg	3	3.53 (2.56-4.50)	3.66 (2.68-4.65)
	4	3.79 (2.82-4.76)	

Table 12: Estimated marginal means for quality of reasoning scores per-test problem and per system for mixed effects models.

Sensitivity Analysis

To show that the results were not sensitive to the design, the inclusion of participants who had used a very early version of Co-Arg, and modification of pre-registered analysis we did three follow-up tests.

First, we evaluated whether there was a gain in quality of reasoning due to repeated testing. We found no evidence to suggest that repeated testing alone could explain improvement in quality of reasoning. We limited our analysis to problems using Google Docs. We compared quality of reasoning between the first and second problem solved (excluding any participants who only worked on one of these problems). Mixed effects linear regression was used with the same design as the prior model, with the exception that data was limited to only reports generated using Google Docs and the independent variable was the problem order (first or second). We found no significant effect of problem order (Coef = 0.15, SE = 0.26, t(47) = 0.58 p = 0.56). On average there was no statistically significant difference in the quality of reasoning score given to the two problems worked on using Google Docs (see Table 12).

Second, we evaluated the main research question using only the data from participants who had no experience with Co-Arg (as outlined as the alternative analysis in the pre-registration). Six students enrolled in the course at CSUSB had used a very early version of Co-Arg in Fall 2017. We evaluated the main research question without these students. This left 45 participants in 11 groups. We found the same pattern of results. We observed a 0.83 increase in quality of reasoning scores from 2.88 to 3.71 out of 10 when participants were using Co-Arg instead of Google Docs (Coef = 0.83, SE = 0.21, t(161) = 3.96, p = 0.0001). Thus, the results are substantively the same whether these individuals with prior experience are included.

Third, we evaluated the effect of our exclusion criteria on the main research question. In the pre-registration we did not state we would exclude participants because they had not submitted a test problem with both systems, to make sure our modification was not affecting our results we repeated our analysis for all users excluding only those who had not completed enough practice. The results were substantially the same (Number of users = 53, Number of Google Docs Reports = 104, Number of Co-Arg Reports = 99, Coef = 0.76, SE = 0.19, df = 149, p < 0.001, Effect Size = 0.48). We then did a further test to show that excluding users due to not enough practice was not affecting results, this analysis included all users and all reports. The results were substantially the same (Number of users = 61, Number of Google Docs Reports = 117, Number of Co-Arg Reports = 104, Coef = 0.83, SE = 0.19, df = 167, p < 0.001, Effect Size = 0.52). Sensitivity analysis suggests that our exclusion criteria had no substantial effect on the pattern of results found and reported.

Research Question 2: *Along which dimensions of quality of reasoning, if any, does Co-Arg with its training improve quality of reasoning on intelligence problems?*

This research question is exploratory and not pre-registered with the Center for Open Science.

We built four linear mixed effects models to test the effect of the system (Google Docs vs. Co-Arg) on each of the four components of quality of reasoning. Each linear mixed effects model used the same design as Research Question 1, that is it tested the fixed effect of System (Google Docs vs. Co-Arg) and included random effects for Problem, Team, and User. Each linear mixed effects model included as its dependent variable a different quality of reasoning dimension subscore (e.g., Hypothesis Generation and Accuracy of

Solution). Each model met the assumptions of the statistical tests with the exception of minor deviations in normality (mixed effect linear regression is robust against minor deviations in normality).

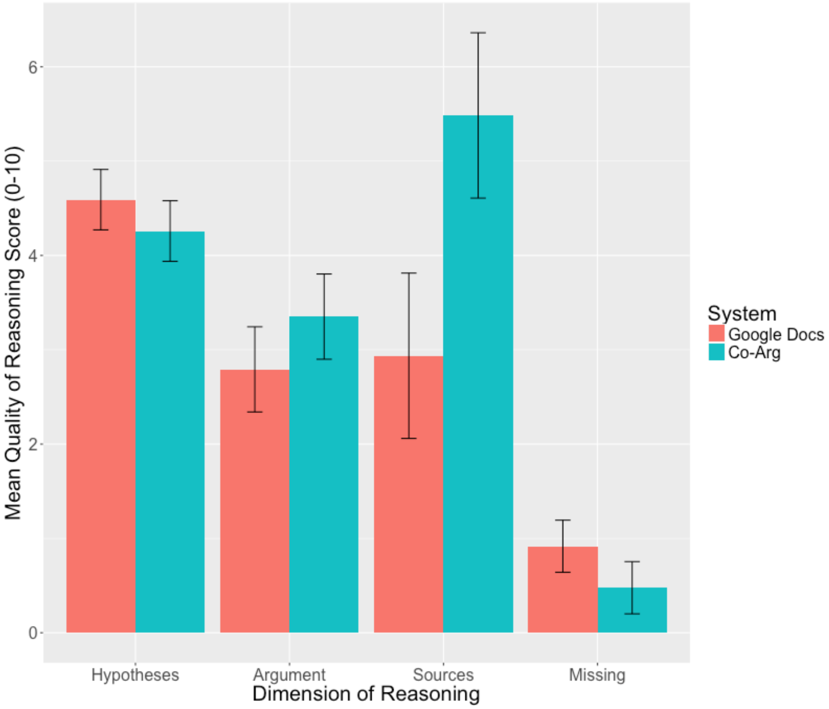


Figure 43: Estimated marginal means and standard errors for each quality of reasoning dimension separated by system usage.

Average scores with Google Docs were low across all four dimensions, with the highest scores for hypothesis generation, followed by evaluation of sources, followed by argumentation, and lastly by evaluation of missing information (see Table 14). The results of the models also showed that Co-Arg helped improve reasoning along two of these four dimensions--argumentation and evaluation of sources-- and had no significant effect on the other two dimensions--hypothesis generation and identification of missing information (see Table 13). We explain the details of each model below in the order of the size of the effect.

Evaluation of Sources of Evidence

Use of Co-Arg was associated with the largest improvement along the dimension of identification of sources and assessment of credibility of evidence. Using Co-Arg was associated with a 2.55 point gain out of 10 along this dimension. On average participants scores improved from 2.94 to 5.48 out of 10 when using Co-Arg as opposed to using Google Docs. This represented a large standardized effect (1.01; 95% CI: 0.75 to 1.26).

Argumentation Structure

In addition, we found that participants improved in their quality of reasoning scores along the dimension of argumentation structure and reasoning. Using Co-Arg was associated with a 0.55 point gain out of 10 along this dimension. On average participants scores improved from 2.79 to 3.35 out of 10 when using

Co-Arg as opposed to using Google Docs. This represented a small standardized effect (0.27; 95% CI: 0.03 to 0.51).

Evaluation of Missing Information

We did not observe a significant difference in quality of reasoning scores along the dimension of identification of key missing information and assumptions. Participants scored slightly higher on identification of key missing information and assumptions when using Google Docs, 0.92 out of 10, compared to when using Co-Arg, 0.48 out of 10. This difference did not reach statistical significance (p = 0.09). Even if a real difference rather than sampling error, this difference represent a small standardized effect (-0.23; 95% CI: -0.49 to 0.04). Using either system, on average participants scored very poorly on evaluation of missing information. This is an area in which substantial improvement could enhance solutions.

Hypothesis Generation

We also did not observe a significant difference in quality of reasoning scores along the dimension of hypothesis generation and accuracy of solution. Participants scored slightly higher on hypothesis generation and accuracy of solution when using Google Docs, 4.59 out of 10, compared to when using Co-Arg, 4.26 out of 10. This difference did not reach statistical significance (p = 0.36). Even if real difference rather than sampling error, this difference represents a very small standardized effect (-0.14; 95% CI: -0.42 to 0.14).

	Hypothesis generation & accuracy of solution		**Argumentation structure & reasoning**		**Identification of sources and assessment of credibility of evidence**		**Identification of key missing information and assumptions**	
	Coef (SE)	p	Coef (SE)	p	Coef (SE)	p	Coef (SE)	p
Intercept	4.59 (0.32)		2.79 (0.45)		2.93 (0.87)		0.92 (0.28)	
System (Co-Arg vs. Google Docs	-0.33 (0.36)	0.36	0.55 (0.25)	0.03	2.55 (0.26)	<0.0001	-0.44 (0.26)	0.09
Standardized Effect Size	-0.13		0.27		1.01		-0.23	

Table 13: Coefficients and standard error results from mixed effects linear regression models evaluating the effect of the system on each dimension of quality of reasoning with user, team, and problem as random effects. System was treated as a factor with Google Docs as the base level. 199 scored reports from 4 problems generated by 51 users in 12 teams.

Marginal Mean (95% CI)				
System	Hypothesis generation & accuracy of solution	Argumentation structure & reasoning	Identification of sources and assessment of credibility of evidence	Identification of key missing information and assumptions
Google Docs	4.59 (3.82-5.37)	2.79 (1.69-3.89)	2.94 (0.26-5.61)	0.92 (0.27-1.56)
Co-Arg	4.26 (3.48-5.03)	3.35 (2.24-4.45)	5.48 (2.81-8.16)	0.48 (-0.17-1.13)

Table 14: Estimated marginal means for each quality of reasoning dimension per system.

Research Question 3: *Do students perceive that their quality of reasoning has improved when using Co-Arg? If so, how?*

This research question is exploratory and not pre-registered with the Center for Open Science.

We found that on average participants perceived Co-Arg to help improve their quality of reasoning. Although there was a trend in which participants felt that Co-Arg helped improve their reasoning more than Google Docs, this difference was not statistically significant. When asked to describe how Co-Arg may have helped to improve their reasoning, participants reported a variety of responses the most frequent of which were evaluating the credibility and relevance of evidence and providing a procedure for thinking about a problem.

Perception of Quality of Reasoning

Participants were asked to separately report whether they thought that Co-Arg and Google Docs had helped to improve their quality of reasoning when working on intelligence problems (“I think that Co-Arg improved my quality of reasoning when solving an intelligence problem (for example, by helping me think about and solve an intelligence problem)”). 47 participants completed this survey (92% completion rate).

System	**Mean**	**95% CI**
Co-Arg	4.06	3.80-4.33
Google Docs	3.77	3.46-4.07

Table 15: Mean scores on a 5 point Bipolar Likert Scale
(1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly agree).

When asked whether Co-Arg helped improve their reasoning participants reported on average 4.06 slightly above Agree on a 5 point bipolar Likert scale (M = 4.06, SD = 0.89, 95% CI = 3.80-4.33). One sample t-test showed that this average rating was significantly greater than a neutral or negative rating on the Likert scale (t(46) = 8.15, $p < 0.001$). When asked whether Google Docs helped improve their reasoning participants reported on average 3.77 between Neutral and Agree on a 5 point bipolar Likert scale (M = 3.77, SD = 1.05, 95% CI = 3.46-4.07). One sample t-test showed that this average rating was significantly greater than a neutral or negative rating on the Likert scale (t(46) = 5.02, $p < 0.001$). When a paired t-test was computed to evaluate whether participants perceived one system to improve quality of reasoning more than the other it failed to reach statistical significance (t(46) = 1.46, p = 0.15). In other words, participants perceived both systems to help improve their quality of reasoning. While there was a trend in which Co-Arg was perceived to help quality of reasoning more than Google Docs, there was insufficient evidence to make this claim.

Participants' Explanations for Improvement in Quality of Reasoning

Participants were asked to explain how Co-Arg improved quality of reasoning (if at all) ("How has your reasoning changed or not changed as a result of working on these intelligence questions with Co-Arg? If your reasoning has changed please explain how it has changed and include concrete examples."). This open ended question was manually coded to extract common themes (see Table 16).

Participants reported a variety of reasons, such as helping them to evaluate evidence, helping to address biases in reasoning, providing a process for reasoning about a problem, making their reasoning more systematic, helping them elaborate their reasoning, and helping to visualize their argument.

The reason stated most frequently was helping students evaluate the credibility and relevance of evidence. This qualitative finding is consistent with the quantitative results measuring actual improvement in quality of reasoning along four dimensions. In terms of scores, we observed the most improvement in evaluation of sources of evidence; this was also a dimension reported by participants in which they felt that they had gained deeper skills and understanding. Participants reported that before using Co-Arg they had not considered evaluating relevance of evidence ("I have learned how much relevance of evidence matters. A piece of evidence may be very credible, but if it is out of date or not really contributing to the exact question asked, I understand that it cannot be effectively used when formulating an argument.", C_T3_P1) and distinguishing credibility from relevance ("it helped me decipher between relevance and credibility" C_T3_P6). They reported that Co-Arg helped them to learn how to judge credibility and reliability ("I didn't understand why some pieces of evidence were more credible or more reliable than others" C_T3_P5). They reported that it helped them incorporate these ratings into their argument ("now I understand why because you have to relate that piece of evidence to other pieces of evidence and see if it fits the theory you have of what happened" C_T3_P5). They reported that it allowed them to base their reasoning more heavily on evidence ("Evidence become more important than ever in my reasoning." B_T3_P3).

Participants also reported that Co-Arg helped them have a procedure to reason about a problem, made their reasoning more systematic, and more detailed. Co-Arg provided participants with a systematic procedure for answering an intelligence question ("My thinking did change in the sense that I now think about defining the hypothesis first then formulate the arguments for and against and using evidence to make connections" C_T5_P2). This helped participants who did not know how to approach the problem or were approaching it using a less systematic method ("My reasoning has become more organized while practicing these intelligence questions." C_T5_P5). Participants also reported that Co-Arg helped them think about a problem slower ("Co-Arg definitely helps a person think slower and helps an individual break down a problem.", B_T3_P1), which led them to be more detailed ("My reasoning has changed working with Co-Arg because it forced me to evaluate every step in detail" B_T4_P4).

Participants felt that Co-Arg helped them make fewer mistakes, by incorporating perspectives from others

("When used properly with input from your team it helps an individual catch something they may have missed or makes one see something from another perspective." B_T3_P1), by teaching them to rely on evidence-based reasoning ("I am now better able to see my own biases and correct them using evidence-based reasoning.", B_T1_P4), by making them be more systematic and detailed in their thinking ("My reasoning has changed working with Co-Arg because it forced me to evaluate every step in detail, reducing the likelihood that I would revert to automatic assumptions in order to streamline my explanation." B_T4_P4), and by making participants more aware of their reasoning ("you can analyze the logic of your argument and forces you to clearly define where you came up with something. thus adding to overall awareness of the line of reasoning so as to not make causal jumps in the future" B_T2_P3).

Participants appreciated having an organized, visual representation of their argument. Participants liked that it represented competing arguments and evidence ("Co-Arg helped me map to multiple competing arguments and visualize the evidence." A_T3_P2). Participants liked having all elements of the argument represented together, which helped them see connections ("Yes Co-Arg made it easier to visualize the problem and made them easier to see how every piece was connected to the main question." C_T4_P3). One participant reported liking the tree format in particular ("Overall, the idea of laying things out in a tree format has been the only real impact.", C_T1_P6).

Theme	Example(s)	Frequency
Evaluating credibility and relevance of evidence	"I think how I evaluate the credibility and relevance of a piece of evidence has drastically changed. I have learned how to decipher between how credible a source can be based on the amount of access and reliable history a source is given. Also, I have learned how much relevance of evidence matters. A piece of evidence may be very credible, but if it is out of date or not really contributing to the exact question asked, I understand that it cannot be effectively used when formulating an argument." [C_T3_P1]	14 (30%)
Helping to address biases in reasoning	"When used properly with input from your team it helps an individual catch something they may have missed or makes one see something from another perspective." [B_T3_P1]	7 (15%)
Providing a procedure for reasoning about a problem	"My thinking did change in the sense that I now think about defining the hypothesis first then formulate the arguments for and against and using evidence to make connections" [C_T5_P2]	7 (15%)
Helping a participant elaborate on their reasoning	"My reasoning has changed working with Co-Arg because it forced me to evaluate every step in detail" [B_T4_P4]	6 (13%)
Making the reasoning process more systematic	"My reasoning has become more organized while practicing these intelligence questions." [C_T5_P5]	6 (13%)
Helping to visualize the argument	"Co-Arg made it easier to visualize the problem and made them easier to see how every piece was connected to the main question" [C_T4_P3]	4 (9%)

Table 16: Most frequent ways participants perceived that Co-Arg helped improve quality of reasoning. 89% of participants out of 47 participants who filled out this question reported that Co-Arg helped improve quality of reasoning on this question.

Research Question 4: *Does Co-Arg with its associated training improve quality of reasoning for some users more than others?*

This research question is exploratory and not pre-registered with the Center for Open Science. We examined a few of moderators to try to understand whether Co-Arg was more beneficial for some types of users than others.

Differences across Courses

Included participants were recruited from courses at three universities: California State University at San Bernardino (CSUSB), University of Omaha (UNO), and George Mason University (GMU). At CSUSB and GMU instructors took an active role in reviewing practice problems with students and helping them learn how to use Co-Arg during the training and practice period. We expected that participants at CSUSB and GMU would learn to use Co-Arg better and thus, we would observed greater improvement when using Co-Arg for these students compared to when using Google Docs. In other words, a significant interaction between a student's course and system use on quality of reasoning scores.

Mixed effects linear regression was computed with quality of reasoning scores as the dependent variable; course, system, and the interaction between the two as independent variables; and team, user, and problem as random effects. We found no significant main effect of a student's course on quality of reasoning scores (see Table 18). We also found no significant interaction between a student's course, the system they were using and quality of reasoning scores. There was a trend in which students from GMU improved the most, followed by students from CSUSB, followed by those from UNO (see Table 17). This trend is in line with our prediction, GMU and CSUSB students show more improvement compared to UNO students. This trend might have reached statistical significance with a larger sample.

Instruction should help participants better understand and use Co-Arg. For the next phase we will examine in more detail how to design instruction to help participants improve their understanding of Co-Arg when completing training and practice.

	Google Docs		**Co-Arg**		**Improvement**
	Mean	95% CI	Mean	95% CI	Mean Difference
CSUSB	3.15	2.10-4.20	3.88	2.83-4.93	0.73
GMU	2.32	1.19-3.46	3.47	2.33-4.60	1.15
UNO	2.95	1.93-3.97	3.61	2.59-4.63	0.66

Table 17: Estimated marginal means for quality of reasoning scores separated by course and system.

	SS	**MS**	**Df (between, within)**	**F**	**p**
System (Co-Arg, Google Docs)	29.8	29.8	1, 147	15.9	<0.001
Course (GMU, UNO, CSUSB)	3.7	1.9	2, 9	1.0	0.41
Interaction between Course and System	1.5	0.8	2,146	0.4	0.67

Table 18: Results of mixed effects linear regression testing the effect of course and the interaction between course and system usage on quality of reasoning scores.

Commonalities among Students who Improved

On average students benefited from using Co-Arg in terms of quality of reasoning, however there was individual variability. Some students showed greater improvement in quality of reasoning scores when using Co-Arg compared to Google Docs than others. To further explore when and how Co-Arg may be helping reasoning we identified a set of students who showed the most improvement and had the highest scores using Co-Arg. We defined improvement as the difference between the average quality of reasoning scores when using Co-Arg and the average quality of reasoning scores when using Google Docs. We took the set of students who had improvement scores in the top 50%. That is the top half of students in our study who showed the most improvement when using Co-Arg. We further limited this set of students to those that had higher average quality of reasoning scores when using Co-Arg, by limiting it to the students who scored in the top 50% of all students when using Co-Arg. This left a set of 19 students that showed both improvement in quality of reasoning when using Co-Arg and better quality of reasoning scores when using Co-Arg. On average these participants scored 4.79 out 10 on quality of reasoning when using Co-Arg (still well below total possible points) and showed average of 2.31 point improvement when using Co-Arg instead of Google Docs (substantially higher than the average gain in quality of reasoning scores).

We then compared this group of participants who benefits the most from Co-Arg to all other participants. Table 19 and Table 20 report the significant differences between the two sets of participants. We observed many commonalities across both groups in terms of demographics, self-reported experience with training, self-reported experience working on the problems, and self-reported usability. The only two significant differences were:

1) Participants who benefited the most from Co-Arg in terms of quality of reasoning reported that they understood the training on evidence based reasoning less than other participants.
2) Participants who benefited the most from Co-Arg in terms of quality of reasoning were more likely to explain the reason as Co-Arg helping them to evaluate the credibility and relevance of evidence.

We also observed marginally significant difference in self-reported understanding following training for Cogent operations and report writing. Similarly, participants who benefited the most from Co-Arg reported less understanding of Cogent operations and report writing.

These results further suggest that Co-Arg helps participants evaluate evidence and better use evidence in generating a solution to an intelligence question. The results are also counterintuitive in terms of the role of training in realizing the benefits of Co-Arg. Understanding evidence-based reasoning, Cogent operations, and report writing are necessary to be able to benefit from Co-Arg, yet we see the opposite pattern of results. This may be explained as a Dunning Kruger effect. Individuals may be bad at assessing their understanding of the training material. Those who only understood the material superficially may have believed they understood it better than they did and self-reported understanding as high. While those who understood the material more, may have realized how complex the concepts were, and self-reported their understanding slightly lower than those who understood the material only superficially. In addition, those participants who reported not understanding the training material as much may have been more motivated to think harder about the material and practice more during the practice and test periods. Future work should use an objective measure of understanding to better gauge the relationship between understanding, perception of understanding, and actual performance.

Category	Variable	Participants who Benefited the most from Co-Arg (Mean)	Other Participants (Mean)	Hypothesis Test	Effect Size Cohen's d
Self-Reported Time	Hours spent training	8.39	7.23	t(17) = -0.73, p = 0.48	0.29
Self-Reported Understanding of Training in	Evidence based reasoning	3.82	4.46	t(16) = 2.35, p = 0.03	-0.96
	Cogent Operations	3.86	4.43	t(17) = 1.97, p = 0.07	-0.79
	Report Writing	3.82	4.23	t(15) = 1.71, p = 0.11	-0.70
	Report Submission	4.13	4.47	t(16) = 1.02, p = 0.32	-0.41
	Analysis Guidance	4.09	4.50	t(15) = 1.30, p = 0.21	-0.54
Self-Reported Experience with Co-Arg	Hours spent using Co-Arg	3.39	3.49	t(32) = 0.13, p = 0.90	-0.04
	Rating as Ideation Tool	4.15	4.02	t(42) = -0.7, p = 0.49	0.20
Self-Reported Usability	SUS Argupedia	56.1	64.0	t(37) = 1.44, p = 0.16	0.43
	SUS Cogent	49.6	52.9	t(42) = 0.69, p = 0.50	0.20
	SUS Co-Arg	52.8	58.4	t(44) = 1.39, p = 0.17	0.39

Table 19: Differences between participants who benefited the most from Co-Arg and all others along quantitative variables.

***Research Question 5:** Does Co-Arg with its associated training improve quality of reasoning for some problems more than others?*

This research question is exploratory and not pre-registered with the Center for Open Science.

Participants worked on four intelligence problems, that differed in terms of the number of hypotheses needed to be evaluated (single vs. multiple hypotheses), the amount of information given, and the difficulty. We expected that Co-Arg might help with some problems more than others. In other words we hypothesized that there would be a significant interaction between problem and system usage on quality of reasoning scores.

We constructed a mixed effects linear regression model with quality of reasoning as the dependent variable; problem, system and the interaction between problem and system as independent variables; and team and user as random effects. We observed a main effect of problem and a marginally significant interaction between problem and system (see Table 21).

Category	Variable	Participants who Benefited the most from Co-Arg(Percent)	Other Participants (Percent)	Hypothesis Test	Effect Size (Cohen's w)
Demographics	Female	41%	29%	$X^2(1) = 0.22$, $p = 0.64$	0.07
	Grad Student	65%	64%	$X^2(1) = 0.00$, $p = 1.00$	< 0.01
	Prior Evidence Based Training	25%	50%	$X^2(1) = 1.40$, $p = 0.24$	0.20
	Related Major	41%	44%	$X^2(1) = 0.00$, $p = 1.00$	< 0.01
Self-Reported Benefit of Co-Arg	Evaluating credibility and relevance of evidence	50%	17%	$X^2(1) = 4.00$, $p < 0.05$	0.53
	Helping to address biases in reasoning	22%	10%	$X^2(1) = 1.05$, $p = 0.31$	0.39
	Providing a procedure for reasoning about a problem	17%	14%	$X^2(1) = 0.06$, $p = 0.80$	0.09
	Helping a participant elaborate on their reasoning	11%	13%	$X^2(1) = 0.06$, $p = 0.80$	0.10
	Making the reasoning process more systematic	6%	17%	$X^2(1) = 1.19$, $p = 0.28$	0.44
	Helping to visualize the argument	11%	7%	$X^2(1) = 0.23$, $p = 0.63$	0.24

Table 20: Differences between participants who benefited the most from Co-Arg and all others along categorical variables.

	SS	MS	Df (between, within)	F	p
System (Co-Arg, Google Docs)	28.4	27.4	1, 150	16.3	<0.0001
Problem	64.6	21.5	3, 150	3.9	<0.0001
Interaction between Problem and System	3.8	3.85	3,42	2.2	0.10

Table 21: Results of mixed effects linear regression testing the effect of the problem, using Co-Arg and the interaction between the problem and using Co-Arg on quality of reasoning scores.

We observed that when using Google Docs users performed the worst on Euclid problem, second worst on the Midland problem, second best on the Mortar problem, and the best on the Bomber problem (Table 22). Although, participants performed better on some problems than others on average participants performed poorly on all four problems. The order of performance on the problems remained similar in Co-Arg as it had in Google Docs, with the exception that Midland and Euclid which switched places in terms of order of performance; users performed the worst on the Midland problem and the second worst on the Euclid problem. In terms of improvement this meant there was the greatest observed improvement for the Euclid problem, the second greatest improvement for the Mortar problem, and the third greatest improvement for the Bomber problem when using Co-Arg. We did not observe improvement for the Midland problem (Table 22).

	Google Docs		Co-Arg		Improvement
	Mean	95% CI	Mean	95% CI	Mean Difference
Midland	2.53	1.85-3.20	2.56	1.88-3.23	0.03
Mortar	3.16	2.50-8.81	4.05	3.39-4.71	0.89
Bomber	3.67	3.01-4.32	4.38	3.72-5.04	0.71
Euclid	2.24	1.58-2.90	3.64	2.97-4.30	1.41

Table 22: Estimated marginal means from mixed effect linear regression model testing the interaction between problem difficulty (as operationalized as single vs. multiple hypothesis intelligence questions) and system usage on quality of reasoning scores.

One interpretation of the results is that there is a trend in which participants improved the most, on average, when using Co-Arg when working on multiple hypothesis questions (Bomber and Euclid) compared to single hypothesis problems (Mortar and Midland). It may be that Co-Arg is more helpful for problems that are more complex and require keeping track of more hypotheses.

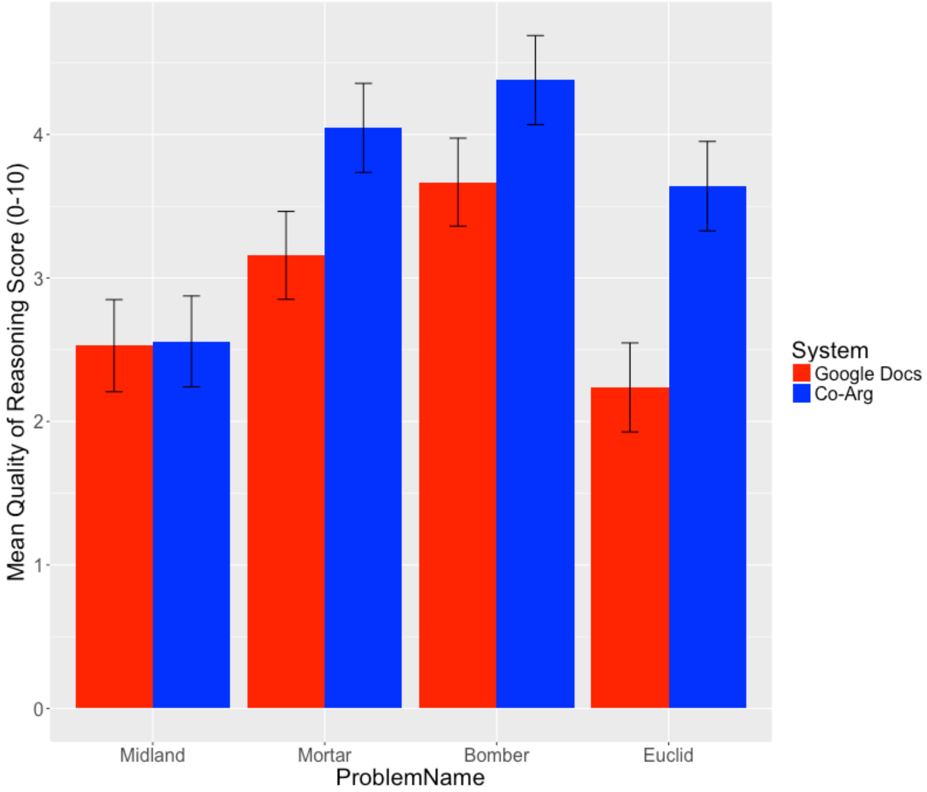


Figure 44: Estimated marginal means and standard errors for quality of reasoning scores across the different problems using each system: Google Docs vs. Co-Arg. Midland and Mortar problems (left) require a single hypothesis, while Bomber and Euclid problems (right) require multiple hypotheses.

7.2.4. Discussion and Conclusions

This study provides empirical evidence to suggest that Co-Arg helps individuals improve their quality of reasoning when answering intelligence questions. We observed a 0.77 point increase out of 10 corresponding to a medium standardized effect size of 0.48 when comparing quality of reasoning when using Co-Arg to quality of reasoning when using Google Docs. Both our assessment of changes in quality of reasoning scores when using the two systems and our assessment of individuals' reported experience when using the two systems suggest that gains in quality of reasoning when using Co-Arg is primarily due to Co-Arg's tools that help users evaluate the credibility and relevance of evidence and incorporate evidence into their reasoning. We also observed, smaller gains along other dimensions. We observed a small improvement in the argumentation structure when using Co-Arg. Users also reported other benefits from Co-Arg, such as that it helped them provide a systematic procedure for reasoning about a problem; be more detailed and less biased when reasoning; and visually represent their reasoning.

We also found that on average even when using Co-Arg participants had poor quality of reasoning scores. With Google Docs on average participants scored 2.89 out of 10 and with Co-Arg on average participants scored 3.66 out of 10. There are a few reasons that participants may have had low scores for quality of reasoning. First, the testing problems were very challenging. Second, the rubric was very detailed and it was very difficult to receive full points. The rubric was designed to be difficult in order to distinguish between fine grained differences in reasoning between products. For example, even if a participant identified the most likely hypothesis they might lose points because they did not identify every other potential hypothesis and/or their probability judgements for some hypotheses were somewhat different from the ideal solution. Third, participants were undergraduate and graduate students, many of whom had no training in intelligence analysis and no interest in intelligence analysis. Thus, compared to the target user, intelligence analysts, these users were less experienced and less motivated. Fourth, using either system participants scored very low in terms of one of the dimensions, identification of key missing information and assumptions. The current version of Co-Arg does not provide tools to assist users with this type of reasoning. It is clear from this study that tools and training are needed to help users with identifying missing information and assumptions.

This study also provided some evidence to suggest that Co-Arg may be more beneficial for some problems than other problems. There was a trend in which Co-Arg helped to improve reasoning the most for Euclid, a multiple hypothesis problem, and did not improve reasoning for Midland, a single hypothesis problem. Co-Arg may be more helpful for more complex problems, such as multiple hypothesis problems, in which a user must keep track of more information. However, differences between single and multiple hypotheses problems cannot explain the results alone. The problems were less difficult than real-world problems--all the information needed to solve a problem was given to participants in a written description, the problems were not open ended, and there was always a clear correct solution. We expect that Co-Arg will be more helpful for more difficult and complex problems.

We also found that our study was limited by the fact that we did not have an objective measure of understanding of Co-Arg and evidence-based reasoning. In this study, we relied on proxies of understanding, such as 1) the amount of practice a user completed using Co-Arg; 2) the amount of instruction a user received during the practice phase in their course, and 3) self-reported understanding after the training, but before the practice. When we related each of these proxies to improvement in quality of reasoning when using Co-Arg we found inconclusive and counter-intuitive results. Amount of practice was not found to be related to improvement in quality of reasoning when using Co-Arg; there was a trend in which more instruction was related to more improvement in quality of reasoning when using Co-Arg, but it was not statistically significant; and greater reported understanding of Co-Arg and evidence based reasoning was associated with less not more improvement in quality of reasoning when using Co-Arg, the opposite of the predicted direction. Because each of these three measures was only a proxy it is difficult to make sense of the results. While it is possible that there is no relationship between understanding of Co-Arg and improvement with Co-Arg; it is more likely that we are not seeing the expected results because each is a poor measure of understanding.

7.3. Study: Assessment of Co-Arg in Improving the Quality of Communication

We found that *using Co-Arg and its associated training improved quality of communication scores on intelligence problems by 1.04 points out of 6.* This improvement in quality of reasoning corresponds with a large effect size of 1.11; above the minimum effect size of 0.1 listed in the BAA for Phase 1. Due to difficulty achieving high inter-rater reliability we propose moving from a holistic measure of quality of communication to one assessing quality of communication along multiple dimensions.

7.3.1. Study Goal and Key Research Questions

This study served as a system evaluation. We evaluated whether Co-Arg and its associated training in evidence-based reasoning improved quality of communication in response to intelligence problems. In this study we investigated one primary research question to evaluate the effect of using Co-Arg on quality of communication. This research question and hypothesis was pre-registered with Center for Open Science; analysis differed slightly from pre-registered plan in two ways:

(1) We altered the number of raters and how they scores were combined (see Section 4.3.2).

(2) To be consistent with Quality of Reasoning analysis we also excluded the 4 reports from users who did not submit at least one report per system (see Section 4.2).

Research Question 1: *Does Co-Arg with its associated training improve quality of communication on intelligence problems for all users?*

Hypothesis 1: *We hypothesized that on average all participants would generate written reports that would score higher on quality of communication in response to post-test problems compared to pre-test problems.*

We expected that Co-Arg would help to improve students' quality of communication by clarifying students thinking allowing them to more easily communicate their conclusion and argument; by providing structure to help compose an organized report; and by helping students to better present evidence as part of their report.

7.3.2. Methods

Data for this study came from the main experiment. We explain the study design, participants, procedures, materials and inclusion criteria in more detail in sections 7.1 and 7.2.

Each written report was scored by three independent raters using the quality of communication rubric from Appendix 9.4. Four possible scores for a report were possible with this 6 point rubric: 0, 2, 4, and 6. Using the rubric raters evaluated a report holistically, assessing multiple aspects of good communication simultaneously, such as main conclusion stated up front, coherent, organized, clear ideas, and justifications provided for conclusions.

Raters had difficulty reaching high inter-rater reliability. In the pre-registration we proposed having only two raters rate each report. However, with only two raters inter-rater reliability was low (ICC = 0.47, 95% CI: 0.31-0.59). To address this issue we added a third rater. With three raters inter-rater reliability reached an acceptable level (ICC = 0.76, 95% CI: 0.70-0.81). The average of the three ratings were taken as our measure of quality of communication for each report.

Statistical analyses were conducted in R using lme4 and lmerTest packages.

7.3.3. Results

Prior to addressing our research question, we examined the relationship between quality of communication and quality of reasoning. Quality of communication scores were moderately, positively correlated with quality of reasoning scores ($r = 0.49$, $t(197) = 7.00$, $p < 0.0001$). The fact that the correlation is not higher suggests that quality of communication is measuring an aspect of the reports that is distinct from quality of reasoning. However, we also observed a moderate correlation suggesting that either 1) students who excel at quality of communication also excel at quality of reasoning or 2) the measures of quality of communication and reasoning are partially overlapping. We proceeded to analyze quality of communication since it was distinct enough from quality of reasoning.

Research Question 1: *Does Co-Arg with its associated training improve quality of communication on intelligence problems for all users?*

We conducted a mixed effects linear regression to evaluate the effect of being trained and using Co-Arg on quality of communication. We used the same statistical model as was proposed in the pre-registration. Quality of communication scores given to reports generated by students were treated as the dependent variable. Which system, Co-Arg or Google Docs, they used to work on the problem was treated as the independent variable of interest. Given the repeated measure design we included random effects in the model. Team was included as a random effect because for all problems students brainstormed as part of a team prior to working on the problem independently. Problem was included as a random effect because we counterbalanced the order in which problems were worked on. User was included as a random effect because each student worked on problems using both systems. The data met all assumptions of the statistical test, with the exception of normality. Deviation in normality of residuals was minimal and mixed effects linear regression is robust against minimal deviations from normality.

We found a significant effect of system usage of quality of communication scores. On average students scored 1.04 points higher on the 6-point measure of quality of communication when using Co-Arg compared to when using Google Docs. On average students scored 2.86 out of 6 on quality of communication when solving intelligence problems with Google Docs and 3.90 out of 6 when solving intelligence problems with Co-Arg. This improvement represents a 1.11 standardized effect size (95% CI: 0.81 to 1.41).

Average quality of communication scores for reports generated using Co-Arg were closest to 4 out 6 on the quality of communication scale. Reports scoring 4 are judged to have conclusions stated up front; to be mostly organized; be clear enough that a reader can easily understand the reasons favoring and disfavoring the main conclusion; however, the analysis supporting these reasons may be difficult to understand; and the description of a few pieces of evidence may be unclear. In comparison, the average quality of communication scores for reports generated using Google Docs were halfway between 4 out of 6 and 2 out of 6. Reports scoring a 2 are judged to have a main conclusion, but it not be stated up front; a lack of coherence and organization; be unclear enough that a reader cannot understand the reasons favoring and disfavoring the main conclusion or the analysis that supports these reasons; and have many items of evidence that are unclear. Thus, the results suggest reports generated using Google Docs are less likely to have a main conclusion stated up front, less organized, more difficult for a reader to ascertain the favoring and disfavoring reasons for the conclusion and analysis, and don't describe the evidence as clearly.

	Coef.	**SE**	**t**	**df**	**p**	**Effect Size**
Intercept	2.86	0.12	24.0	11.0	<0.0001	
System (Co-Arg vs. Google Docs)	1.04	0.12	8.3	146.3	0.0001	1.11

Table 23: Results of mixed effects linear regression model evaluating the effect of the system on quality of communication with user, team, and problem as random effects. System was treated as a factor with Google Docs as the base level. 199 scored reports from 4 problems generated by 51 users in 12 teams. Variance for random effects were user = 0.06, team = 0.04, problem = 0.01, error = 0.77.

System	**Problem Order**	**Marginal Mean (95% CI)**	**Marginal Mean (95% CI)**
Google Docs	1	2.98 (2.68-3.29)	2.86 (2.60-3.12)
	2	2.75 (2.45-3.05)	
Co-Arg	3	3.91 (3.60-4.22)	3.90 (3.64-4.16)
	4	3.89 (2.58-4.20)	

Table 24: Estimated marginal means for quality of communication scores per test problem and per system based on results of mixed effects model.

7.3.4. Discussion and Conclusions

This study provides evidence to demonstrate that using Co-Arg helped users to improve how they wrote up their solutions. *On average users reports scored 1.04 points higher on a 6 point scale when using Co-Arg compared to Google Docs, which represents a large sized effect of 1.11; above the minimum effect size of 0.1 listed in the BAA for Phase 1.* When using Co-Arg, on average reports had a main conclusion stated up front, were mostly coherent and organized, and had several clear ideas. In comparison when using Google Docs reports were on average less likely to have a main conclusion stated up front, were less likely to be well organized, did not make the reasons favoring and disfavoring the main conclusion clear enough, and did not explain the evidence as well.

While we were able to establish high enough inter-rater reliability for quality of communication when three raters rated each report, raters experienced a great deal of difficulty using the holistic rubric. This meant that inter-rater reliability was low with only two raters. Raters struggled to come to agreement even during calibration meetings meant to work through disagreements in the application of the rubric to reports. In addition, the quality of communication rubric collapsed several dimensions into one rating. There are multiple aspects of quality of communication that should be measured separately and if measured separately may make it easier for raters to assess a report and to understand the aspects of communication that Co-Arg is helping the most.

7.4. Study: Usability of Co-Arg

We found that Co-Arg is at least moderately easy to use, scoring 56.3 on average on the system usability scale (SUS), slightly above the hypothesized score of 55.

7.4.1. Study Goal and Key Research Question

This study served as a system evaluation. In this study we investigated the research question on whether the participants find Co-Ag easy to use; this research question and hypothesis was pre-registered with the Center for Open Science analysis did not differ from proposed plan in the pre-registration.

Research Question 3: Do participants find Co-Arg easy to use?

Hypothesis 3: We hypothesize that participants will on average rate Co-Arg as being at least moderately easy to use, that is scoring at least 55 on average on the SUS.

Follow-up analyses (e.g. NPS, separate SUS scores by component) were exploratory and not pre-registered.

7.4.2. Method

Data for this study came from the main experiment; the design, participants, procedures and inclusion criteria are explained in more detail in Section 4.1. We evaluated the usability of Co-Arg by using the System Usability Scale (SUS). SUS is a standard 10-item scale questionnaire that provides an overview of subjective assessments of usability (see Appendix 9.2). SUS yields a single number representing a composite measure of the overall usability of a system. 47 participants who were included answered the questionnaire. Participants assessed Cogent and Argupedia separately. We determined to what extent Co-Arg was easy to use by computed the average scores for Argupedia and Cogent.

7.4.3. Results

Figure 45 shows the average SUS scores for Argupedia, Cogent, and Co-Arg, separately for the users at the three universities, and overall for all users.

Co-Arg scored 56.3 (95% CI: 52.0-60.6) on average, slightly above the hypothesized score of 55. However, this difference did not reach statistical significance. A one sample t-test showed that SUS scores were no different from 55 (t(46) = 0.62, p = 0.54, Cohen’s d = 0.09). Nonetheless a score of 56.3 with 95% confidence interval between 52.0 and 60.6 suggest that participants considered the system as reasonably easy to use.

The evaluation of Argupedia resulted in a 61.0 SUS score and Cogent received 51.7 Considering the mapping to adjective scales proposed by Bangor et al. (2009), these scores can be interpreted as ‘good’ and ‘ok’ respectively.

The SUS scores varied across the users at the three universities: Argupedia received the highest score at the University of Omaha (score of 66.0), the next highest score at George Mason University (61.1), and the lowest score in the experiment conducted at San Bernardino (score of 52.6). Cogent received the highest score at GMU (score of 56.9), the next highest score at UNO (53.3), and the lowest score (44.4) at CSUSB. In general, Argupedia received higher scores than Cogent. We hypothesize that this is due to the specific tasks executed in each of these two components.

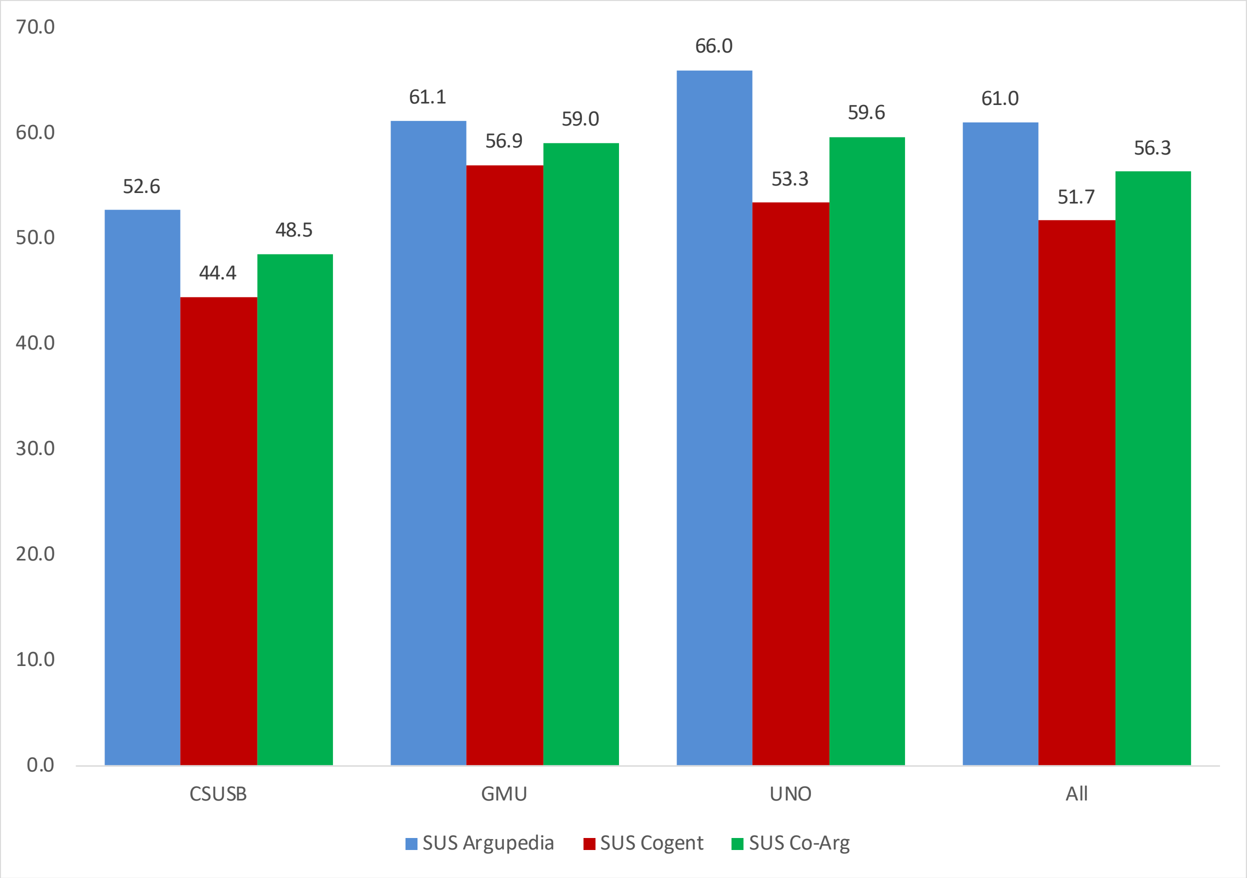


Figure 45: Average SUS scores for Argupedia, Cogent and Co-Arg.

We also assessed Co-Arg and its components by using the Net Promoter Score (NPS). Largely adopted among business managers, NPS is suitable for commercial products, as an effective measure to predict a business growth. The Net Promoter Score is a "customer loyalty metric" based on the responses of a single question: "How likely is it that you would recommend our system to a friend or colleague?" The scoring for this answer is most often based on a 0 to 10 scale (see Appendix 9.3).

Figure 46 shows the results obtained by Co-Arg, Argupedia, and Cogent. Here, as opposed to the SUS scale, Cogent received higher scores than Argupedia.

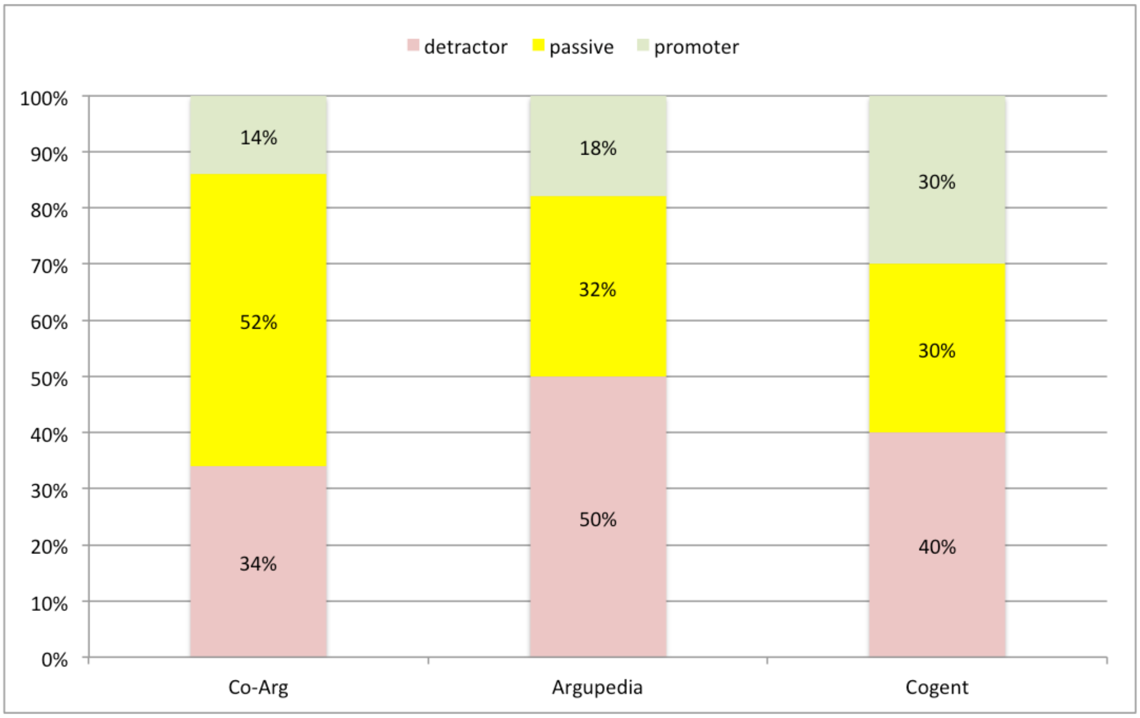


Figure 46: Net Promoter Scores for Co-Arg, Argupedia, and Cogent.

7.4.4. Discussion and Conclusions

This study provides evidence to demonstrate that using Co-Arg is at least moderately easy to use, scoring 56.3 on average on SUS.

Argupedia received higher SUS scores probably because the majority of the study participants (characterized as young adults, college students) are familiar, and therefore more comfortable to interact, with a web-based interface. In addition to that, the nature of the tasks executed in Argupedia, and their social component (collaboration and communication with peers) may also have influenced the participants' perceptions of "ease of use."

Cogent does require participants to follow a training before they are able to execute the activities supported by the system. Therefore, there is an inherent learning curve for users, which added to the fact that most of them were not familiar with intelligence analysis.

7.5. Study: Evaluating Training in Evidence-based Reasoning with Co-Arg

We explored how much time and effort was self-reported to have been spent on training and practice; whether users reported understanding key concepts and operations; and whether better understanding was associated with higher ratings of usability for Co-Arg. We found that on average 29 hours would have been spent on training and practice if users had done all 5 practice problems, but that in reality, users only completed two and half practice problems. Self-reported understanding of material was high, but free-response comments suggested that specific evidence-based reasoning concepts were difficult for some users. In the future we plan to assess understanding using an objective measure. Users described the complexity of evidence-based reasoning and report development. They valued the organization of online training modules for later reference, and they suggested that practice embedded within the training videos could improve their proficiency in building argumentation and producing reports.

All the research questions in this section are exploratory and not pre-registered with the Center for Open Science.

7.5.1. Study Goal and Key Research Questions

The overarching goal of this study was to evaluate the training and practice materials given to participants to teach them how to use Co-Arg and evidence-based reasoning. We evaluated the training and practice in several different ways. First, we assessed how much time was spent on training and practice and the degree to which participants completed the assigned practice. We asked two research questions:

1. How much time was self-reported to have been spent on training and practice?
2. How much practice was completed?

By assessing how much time was reported to have been spent we can partially evaluate how time consuming the assigned training and practice was. Participants cannot benefit from practice they don't complete. By assessing the amount of practice participants complete we can assess participants' willingness to engage with the materials for as much time as suggested.

Second, we assessed whether participants reported understanding the training after they watched the videos and completed the tutorials. We expected that participants would report understanding the material. However, if there were particular concepts or operations they did not understand, we wanted to identify these. We asked two additional research questions:

3. Did participants report understanding the training material?
4. Were there any operations or concepts that participants reported not understanding?

Third, we assessed whether completing and understanding the training and practice helped improve the usability of Co-Arg. We expected that Co-Arg would be easier to use with adequate training and practice, we wanted to assess this prediction. We asked three additional research questions:

5. Is more time spent on training and practice associated with greater usability?
6. Is completing more practice associated with greater usability?
7. Is better self-reported understanding of concepts needed to use Co-Arg associated with greater usability?

Fourth, we assessed participants' suggestions for how to improve the training and practice, and asked two additional research questions:

8. How do participants suggest improving the training?
9. How do participants suggest improving the practice?

7.5.2. Methods

Data for this study came from the main experiment; the design, participants, procedures and inclusion criteria are explained in more detail in Section 4.1. Participants completed a self-report questionnaire at the end of the training and at the end of each practice problem. The training questionnaire asked participants to report the number of hours spent on the five training modules: Evidence Based Reasoning, Cogent Operations, Report Writing, Report Submission, and Analysis Guidance. In addition, participants were asked to report the degree to which they understood key concepts and operations associated with these five modules ("Please mark the degree to which you agree or disagree that you have understood each of the following concepts"), concepts such as hypothesis, relevance, AND arguments, and operations, such as define an item of evidence and move a subtree under a hypothesis. Multiple items were given per module, each was rated on a 5 point Likert Scale from 1 = Strongly disagree to 5 = Strongly agree. Participants were also asked to make suggestions about ways to improve the training ("Did you experience any difficulties with these hands on exercises? Please explain" "Do you have any suggestions on how to improve this training? Please elaborate.").

Participants were asked to complete a practice questionnaire following each of the five practice problems. The practice questionnaire asked participants to report the number of hours they spent practicing. It also asked participants to suggest ways to improve the practice ("How can we improve the practice? Do you have any suggestions?").

Participants were assigned to complete 5 practice problems. Some participants did not complete all 5 problems, others partially completed some or all of the problems. Reports submitted for practice problems were scored for completion. A practice report was considered complete if the participant submitted a report with a formal argument, a probability for the top hypothesis, and independent work in Cogent. A practice report was considered partially complete at a medium level if they provided a formal argument and a probability of a top hypothesis, but no evidence of independent work in Cogent. A practice report was considered partially complete at a low level if they submitted a report but there was no formal argument. A practice problem was considered incomplete if no practice report was submitted or no solution with a probability of a top hypothesis was submitted.

For the purposes of this study we included all participants, since inclusion criteria depended in part on how much practice was completed. 35 out of 62 participants completed the training questionnaire (56% completion rate). 44 out of 62 participants completed at least one practice questionnaire (71% completion rate). On average participants completed the practice questionnaire for 2.77 practice problems.

7.5.3. Results

Research Question 1*: How much time was self-reported to have been spent on training and practice?*

This research question is exploratory and not pre-registered with the Center for Open Science.

Participants reported spending an average of 8.40 hours on training and 4.13 hours per practice problem. If a participant completed all 5 practice problems this would result in a total of 29.05 hours of training and practice on average if self-reported times are accurate.

	Practice & Training	Hours Mean (95% CI)
Training Component	Total	8.40 (6.92-9.88)
	Evidence Based Reasoning	4.26 (3.33-5.20)
	Getting Started Cogent	1.09 (0.80-1.37)
	Cogent Operations	0.62 (0.44-0.79)
	Report Writing	1.44 (1.15-1.73)
	Automatic Report Generation	0.54 (0.34-0.75)
	Conducting Analysis	0.51 (0.34-0.68)
Practice	Average per Problem	4.13 (3.72-4.53)

Table 25: Self-reported mean hours spent on each component for the training and practice. Cogent module is split into Getting Started Cogent with Cogent and Cogent operations.

Research Question 2: *How much practice was completed?*

This research question is exploratory and not pre-registered with the Center for Open Science.

Participants engaged in some practice, but many did not perform as much as was assigned. Only around half of the practice problems were fully completed. A substantial percentage of practice problems were only partially completed and or not completed at all. On average participants completed 3.87 problems at least at low level (95% CI: 3.45-4.29). On average participants completed 3.16 problems at least at a medium level (95% CI: 2.63-3.69). On average participants fully completed 2.4 problems (95% CI: 1.9-2.9).

Not Completed	Partially completed (very little work - Low)	Partially completed (moderate amount of work - Medium)	Completed
70 (23%)	44 (14%)	47 (15%)	149 (48%)

Table 26: Frequency and percent of practice problems completed, partially completed, or not completed. Participants were assigned five practice problems per person.

Research Question 3: *Did participants report understanding the training material?*

This research question is exploratory and not pre-registered with the Center for Open Science.

Participants were given training videos and hands on tutorials on different aspects necessary for using Co-Arg. Participants then rated the degree to which they understood the corresponding aspects of Co-Arg using 5 point Likert scales. Items corresponding to a single component were averaged to produce a measure of understanding per component. All scales achieved high internal consistency (Cronbach alpha > 0.90) suggesting that they could be treated as a single scale per module.

Across the five modules participants on average reported scores between 4 and 5 which fell between Agreeing and Strongly Agreeing that they understood the material and operations (see Table 27). One sample t-tests showed for each tutorial that the score given was significantly greater than Neutral, suggesting that on average participants did report understanding the material.

Co-Arg Tutorial Component	Self-Reported Understanding Mean (95% CI)	Hypothesis Test
Evidence Based Reasoning	4.19 (3.94 - 4.43)	t(34) = 9.78, p < 0.001
Cogent Operations	4.18 (3.92 - 4.44)	t(34) = 9.35, p < 0.001
Report Writing	4.06 (3.85 - 4.28)	t(34) = 10.02, p < 0.001
Automatic Report Generation	4.34 (4.06 - 4.61)	t(34) = 9.91, p < 0.001
Conducting Analysis	4.33 (4.07 - 4.58)	t(34) = 10.54, p < 0.001

Table 27: Participants reported the degree to which they understood material taught in the different training videos using a 5 point Likert scale where 1 = Strongly disagree, 3 = Neither agree or disagree, 5 = Strongly agree.

Research Question 4: *Were there any operations or concepts that participants reported not understanding?*

This research question is exploratory and not pre-registered with the Center for Open Science.

For each tutorial participants reported their understanding for key concepts and operations. Although we found that overall participants reported understanding the tutorials, we wanted to evaluate if there were any concepts or operations that they did not understand as well. We recorded the items for which participants on average rated that they understood the concept or operation less than Agree (4 on the 5 point Likert scale). Participants had average scores less than 4 for only 5 out of the 45 items (11%). Three of these items were about evidence-based reasoning. On average participants reported a score of 3.89 (95% CI: 3.54-4.23) for how well they understood inferential force, 3.91 (95% CI 3.59-4.24) for how well they understood AND arguments, and 3.91 (95% CI: 3.70-4.12) for how well they understood on balance arguments. Participants also reported on average a score of 3.96 (95% CI: 3.60-4.31) for how well they understood how to move a subtree under a hypothesis. Participants reported on average a score of 3.97 (95% CI: 3.73-4.21) for how well they understood how to add argumentation fragments to the report. Each of these items average reported understanding was only slightly below agree and significantly greater than neutral. However, these are areas where improvement in training may help understanding.

In their survey responses, participants were asked to elaborate on low ratings for understanding. Their comments were consistent with their quantitative responses. One user suggested that “there was too much information to comprehend all at once”, while another user claimed that it was “a lot to take in one week.” These responses suggest that users could benefit from follow-on training, opportunities to ask questions, and scaffolded supports for challenging concepts and operations.

Research Questions 5-7: *Is more time spent on training and practice associated with greater usability? Is completing more practice associated with greater usability? Is better self-reported understanding of concepts needed to use Co-Arg associated with greater usability?*

These research questions are exploratory and not pre-registered with the Center for Open Science.

In general, training and practice behavior and self-reports were unrelated to usability scores (see Table 28). There were two exceptions. First, there was a marginally significant correlation between self-reported understanding of evidence based reasoning and the reported usability of Argupedia. Participants reported finding Argupedia easier to use if they also reported understanding evidence based reasoning concepts. This makes sense, since the design of Argupedia tasks are based in part on evidence-based reasoning.

Second, there was a marginally significant correlation between self-reported understanding of how to compose a report and the reported usability of Argupedia. Participants who reported understanding how to compose a report also reported finding Argupedia easier to use. Although Argupedia was not used to generate reports, participants who better understood the elements that needed to go into a report may have found the functions of Argupedia, such as hypothesis generation and rating evidence, more intuitive.

Training and Practice Component		**SUS Argupedia**	**SUS Cogent**
Training	Training Total Hours Spent	r = -0.14, df = 28, p = 0.45	r = -0.12, df = 28, p = 0.52
Self-reported Understanding Training	Evidence Based Reasoning	r = 0.33, df = 31, p = 0.06	r = 0.21, df = 31, p = 0.23
	Cogent Operations	r = 0.23, df = 31, p = 0.18	r = 0.21, df = 31, p = 0.25
	Report Writing	r = 0.32, df = 31, p = 0.07	r = 0.09, df = 31, p = 0.61
	Automatic Report Generation	r = 0.21, df = 31, p = 0.23	r = 0.11, df = 31, p = 0.52
	Conducting Analysis	r = 0.22, df = 31, p = 0.20	r = 0.17, df = 31, p = 0.17
Practice	Practice Average Hours per Problem	r = 0.14, df = 99, p = 0.16	r = 0.03, df = 99, p = 0.74
	Number of Practice Problems Completed	r = 0.02, df = 48, p = 0.89	r = -0.19, df = 48, p = 0.18

Table 28: Correlations between amount of time spent on training and practice, amount of practice completed, and self-reported understanding of Co-Arg components and self-reported usability of Co-Arg, as measured by ratings of Argupedia and Cogent using the System Usability Scale.

Research Questions 8-9: *How do participants suggest improving the training? How do participants suggest improving the practice?*

Both these research questions are exploratory and not pre-registered with the Center for Open Science.

The training and the practice were offered in a linear sequence. Users viewed 1-4 minute videos organized by topic prior to engaging with Co-Arg on practice problems. Multiple users suggested that an integrated format in which they engaged in hands-on practice on a problem in parallel with viewing videos would improve their learning: *I feel if you were allowed to access Cogent during those video lessons it would be easier to follow along with some of the concepts and work alongside them, instead of having to go back and having to watch the videos again after gaining access to Cogent during the hands-on section.*

Participants offered mixed perspectives on the modular training format. Some users commented on the necessity to "refer back" to specific videos as they build their argumentation and generated their reports. The modular format supported users in selecting specific videos for reviewing. Others reported that the module format increased the duration and complexity of the training, as the students were required to click on multiple short-length videos in sequence.

Finally, participant comments suggested a need to revisit assessment of understanding during the training. The multiple-choice quizzes with retake options were designed to provide self-assessment of knowledge from preceding videos, the quizzes were "paced nicely with the lecture." One participant commented that their format was unwieldy because they could not focus only on questions they had missed, while others found the questions confusing. Another participant commented on the potential advantage of practice-based assessment with complex material in addition to the quizzes: *It is a long* *training process, so practicing and getting feedback during the training will help ensure that all the information is being absorbed.*

7.5.4. Discussion and Conclusions

A substantial amount of time was reported to have been spent on training and practice. On average 8 hours were spent on training and 4 hours per practice problem. For participants who completed all 5 assigned practice problems this would have totaled 29 hours of training and practice. However, we found that most participants did not fully complete the practice problems assigned. On average, participants fully completed only 2.4 practice problems. The amount of training and practice given to users in this study may have been more than they were willing to engage with. In addition, there may not have been enough instruction, feedback, and scaffolding to motivate users to complete the practice.

We should note that this training was used in the pilot experiment conducted by T&E in Spring 2018, when none of the participants completed it. As a result, we significantly improved the Argupedia module that controls the training, to make the training more adaptable to various user preferences and needs. However, we were not able to use this improved training in our internal experiments because they were already started.

After training, a majority of users reported understanding evidence-based reasoning and Co-Arg operations. On average users reported between agreeing and strongly agreeing that they understood each of the modules. This suggests that the training and practice was at least partially useful in helping users understand key concepts and operations. We were also able to identify a few topics that users understood less.

We did not find any significant relationships between training and practice on one hand, and usability on the other hand. There was a marginally significant correlation between self-reported understanding of evidence-based reasoning and usability of Argupedia. We expect that users who understood evidence-based reasoning and operations would find it easier to use Co-Arg. The proxies we used for understanding, such as amount of practice and self-reported understanding may not be good gauges of understanding. Future work will examine whether an objective measure of understanding is related to higher usability. If we do find a relationship between the two, it may in part help us improve usability scores by providing users with the training that makes Co-Arg more intuitive and easier to use.

7.6. Study: Evaluation Grid for Quality of Reasoning

Evaluating intelligence reports for evidence of higher order thinking skills is a major challenge because:

- There may be more than acceptable solution for a problem, especially if tightly-banded probability assessments are included as part of the solution
- Dimensions for quality of reasoning are very interrelated and thus hard to break out for separate assessments (e.g., reasoning should simultaneously lay out the credibility of evidence, reasoning supporting the judgment and reasoning supporting the probability assessment)

- Participants in the testing are likely to use language that is not identical to language in the evaluation rubric, and raters must then judge whether the language is consistent with the meaning of the language in the rubric
- The communication skills of participants vary, and the communication skills of some participants may not be commensurate with their reasoning skills
- Not all elements will apply for each problem—especially at the simple and slightly challenging levels
- Logical errors can take multiple forms and cannot necessarily be identified in advance
- Products likely to exhibit mixed performance against criteria

We experimented with two different types of evaluation grids. The results obtained are described below.

7.6.1. Evaluation Grid with General Scoring Guidance

In the Fall 2017 experiment with Co-Arg at CSUSB we used a QoR grid that attempted to achieve the right balance between common standards of quality of reasoning and problem-specific expected reasoning (see Appendix 6.1). This grid used the six criteria listed below. Inter-rater reliability, as measured by intraclass correlation coefficients, is provided next to each criterion.

- Accuracy of solution (ICC = 0.79, 95% CI: 0.49-0.92)
- Argument structure and reasoning (ICC = 0.65, 95% CI: 0.00-0.87)
- Evidence assessment (ICC = 0.56, 95% CI: -0.09-0.82)
- Uncertainty (ICC = 0.46, 95% CI: -0.22-0.77)
- Assumptions (ICC = 0.09, 95% CI: -0.88-0.61)
- Analysis of alternatives (ICC = 0.36, 95% CI: -0.66-0.75)

The experimental results showed poor inter-rater reliability for the last four of the above six criteria. To address this weakness, we re-organized the QoR criteria as follow:

- Combined “Accuracy of solution” and “Analysis of alternatives” into a single criterion
- Eliminated “Uncertainty” because it is a characteristic of each of the other criteria

The new four criteria are:

- Hypothesis generation and accuracy of solution
- Argument structure and reasoning
- Identification of sources and assessment of credibility of evidence
- Identification of key missing information and assumptions

We hypothesize that this streamlined grid will produce better IRR results, but we have not tested this hypothesis because IARPA has opted in favor of a problem-specific grid for the T&E evaluation and we wanted to use the same type of grid in our internal evaluations.

7.6.2. Evaluation Grid with Problem-Specific Scoring Guidance

This grid used in our internal testing is very similar to the grid used for the T&E experimentation. Appendix 9.6 provides an example of such a grid. Notice that it uses the same four criteria and the refinement of the previous grid.

To further improve the chances of obtaining high IRR, we organized workshops with all the raters to calibrate their evaluations for each of the problems used in the internal experiments. For each distinct problem, each rater was given a sample report to evaluate and then the resulting evaluation scores were discussed to understand the differences and improve the clarity of the evaluation grid. As a result, we obtained a high inter-rater reliability for all the criteria.

Using this new evaluation grid in the Spring/Summer 2018 internal evaluation we were able to achieve high inter-rater reliability for the composite measure of quality of reasoning (ICC = 0.85, 95% CI: 0.80-0.88), and good enough inter-rater reliability for each criterion:

- Hypothesis generation and accuracy of solution (ICC = 0.85, 95% CI: 0.80-0.88)
- Argument structure and reasoning (ICC = 0.82, 95% CI: 0.77-0.87)
- Identification of sources and assessment of credibility of evidence (ICC = 0.82, 95% CI: 0.77-0.86)
- Identification of key missing information and assumptions (ICC = 0.63, 95% CI: 0.52-0.72)

7.6.3. Conclusions

The four criteria for quality of reasoning was a good way to separate different aspects of quality of reasoning.

The problem-specific scoring guidance worked well because the problems were constrained-enough to have a unique solution. For less-constrained problems, with many possible solutions, a general scoring guidance seems more appropriate.

In any case, workshops with the raters to calibrate their evaluations for each problem is necessary.

8. Use of Co-Arg in the T&E Experiment

8.1. Participation Numbers

T&E created 8621 user accounts for the evaluation. Out of these, 368 were reserved for the individual control group in the standard condition, evenly spread over 4 blocks, and 8253 user accounts were allocated to the choice condition. 138 out of the 8253 choice user accounts were distributed to IARPA, T&E and performer users for evaluation monitoring and system testing, with the remaining 8115 accounts available to the actual participants in the evaluation.

The user accounts were created in 3 stages:

- 1792 accounts were created one week before the experiment started, when the evaluated systems where brought up together with the T&E portal, for users that consented to the experiment and completed the initial surveys
- 3920 accounts were created right before the experiment started (airdrop 1) for 1248 users that consented to the experiment and (partially) completed the initial surveys after the first round of

accounts were created, and 2672 users that consented to the experiment but did not work at all on the surveys

- 2909 accounts were created 3 weeks into the experiment, after the training and practice phases, and right before the challenge phase; 461 of the accounts were created for users that consented and (partially) completed the initial surveys, while 2448 accounts were created for users that only consented to the experiment

During the training phase (the first week of the experiment), users could train with as many of the performer systems as they wanted. The 4 performers were asked to provide a 5-minute promotional video for their systems that experiment participants could view before choosing the systems with which to train. The number of users that clicked on the promotional videos, as recorded by the T&E portal and shown in Table 29, reveals an almost even distribution between the performers. In particular, the Co-Arg video was viewed by 129 users.

	Started
bard	142
coarg	129
swarm	124
trace	136

Table 29: Video viewing (from data provided by T&E on August 29, 2018).

Training participation was monitored at two stages. First, the main experiment portal managed by T&E recorded when users accessed an evaluated system for the first time during the training week and counted that as the users starting training. Then, each evaluated system tracked user training activities for that system and reported back to the T&E portal when a user completed the corresponding training.

The Co-Arg system provided comprehensive training material that included video and text instruction followed by hands-on activities with Cogent. Due to the importance we placed on having properly trained users, we configured Co-Arg to require them to complete the training before they could solve any problem with the system, even after the official training period ended. For example, if a user decided to use Co-Arg to solve a Round 3 problem, they were required to complete the training at that time if they did not do so previously, before they were allowed to start working on that problem. Therefore, the training data collected by T&E very close to the end of experiment offers a more complete picture of the training activities with Co-Arg.

As shown in Table 30, a total number of 586 users performed some training with Co-Arg, comparable with the other evaluated systems. 284 of those users completed the Co-Arg training. The relatively low number of training completions with Co-Arg compared to the other evaluated systems is explained by the much more comprehensive training required by Co-Arg, which required more commitment than some of the users were able to make.

	Completed	In-progress	*Total*
bard	282	12	*294*
coarg	284	302	*586*
swarm	510	91	*601*
trace	637	33	*670*
ctrl_team	469	2	*471*
ctrl_ind	567	3	*570*
Total	*2749*	*443*	*3192*

Table 30: Training Participation (from data provided by T&E on August 29, 2018).

The training period was followed by a two-week practice period in which users could solve up to 2 problems with each evaluated system in conditions similar to the challenge phase that would follow, to get more hands-on practice. The Co-Arg team provided 2 problems to be solved with the system during this period: Salazar and Manada SAM Sale. As shown in Table 31, 205 users worked on the Manada SAM Sale problem, and 157 users worked on the Salazar problem, for a total number of 362 users, comparable with the other evaluated systems. A number of users worked on both of the Co-Arg practice problems.

During the following 3 rounds of challenge problem solving, the number of users involved with each evaluated system decreased significantly, as shown in Table 31. In particular, 117 Co-Arg users were active during Round 1, 53 during Round 2 and 41 during Round 3, comparable with the number of users active with the other evaluated systems. An encouraging fact is that Co-Arg users were relatively evenly distributed among the 4 challenge problems offered in each round, even though some of the problems were not the best suited for Co-Arg's approach to evidence-based reasoning (for example those involving the construction of a Bayesian network).

The challenge phase was designed by T&E in such a way that required users to commit to only one system for solving each problem. Once a user selected the system for solving a problem, they were not allowed to switch to a different system for that problem. A few users chose to solve all challenge problems exclusively with Co-Arg.

For the T&E experiment, Co-Arg employed a dynamic team formation strategy that took into account the desired team size and the distribution of the work schedule between Argupedia and Cogent. There was no control over which users will choose to work with Co-Arg on which problems, or when will they start working on them. Because the Co-Arg work schedule reserved a percentage of time to collaboration between team members in Argupedia followed by individual work in Cogent, it was important to maximize the availability of team members during the work scheduled in Argupedia. Adding a new user to a team late in the work schedule meant that the team would not benefit from the new user's contributions much because the existing members were most likely working individually in Cogent by then. Also, new users would find an already developed analysis that would frame their thinking process and force them to work more in Cogent to adapt it to their own solution as they would not be able to change it much in Argupedia.

		Control	Choice						
		ctrl_ind	ctrl_ind	ctrl_team	trace	bard	coarg	swarm	*Total*
Practice	pulse	62	131	65	.	.	.	.	*258*
	rakona	46	83	41	162	.	.	.	*332*
	cellnet	.	.	.	.	203	.	147	*350*
	kalukistan_bomb	.	.	.	244	.	.	148	*392*
	manada_sam_sale	.	.	.	.	.	205	.	*205*
	smoking_and_cancer	.	.	.	.	223	.	.	*223*
	salazar	.	.	.	.	.	157	.	*157*
	Total	*108*	*214*	*106*	*406*	*426*	*362*	*295*	*1917*
Round 1	black_site_surveillance	8	12	8	34	74	25	39	*200*
	missile-test	17	18	5	37	36	31	24	*168*
	lifehacker	14	16	11	43	23	22	32	*161*
	arthur_allen	14	26	10	83	25	39	54	*251*
	Total	*53*	*72*	*34*	*197*	*158*	*117*	*149*	*780*
Round 2	cyberattack	12	12	4	12	35	12	15	*102*
	zest	14	10	4	17	7	8	13	*73*
	which_lovell	5	11	2	20	6	12	17	*73*
	who-is-the-spy	11	17	8	33	10	21	39	*139*
	Total	*42*	*50*	*18*	*82*	*58*	*53*	*84*	*387*
Round 3	frogs	4	4	4	19	11	9	11	*62*
	without_a_trace	14	8	5	24	10	9	16	*86*
	dark_web	6	9	4	27	11	13	19	*89*
	prison_break	8	10	5	28	6	10	23	*90*
	Total	*32*	*31*	*18*	*98*	*38*	*41*	*69*	*327*

Table 31: Problem Solving Participation (from data provided by T&E on August 27, 2018).

Therefore, Co-Arg imposed some time limits for adding users to existing teams versus creating new teams for the new users. If the desired team size was reached before the first time limit, or if the minimum team size was reached before the second time limit, new teams were created for new users. As a result, more teams were dynamically created by Co-Arg during all the phases of the experiment than by the other evaluated systems, as shown in Table 32 (there was no team creation during the training week, as Co-Arg users trained individually).

47 teams were created by Co-Arg during the practice period, compared to 9 by the control/team system (Concur), 21 by BARD and 14 by SWARM. A similar situation was encountered during the challenge rounds, where Co-Arg created in each round more teams than all the other evaluated team-based systems combined. A consequence of this situation is that the average number of users per team was lower (and in some cases significantly so) for Co-Arg than for the other evaluated systems.

For the choice condition in the T&E experiment Co-Arg was configured to allow any team member who so desired to personalize the team analysis in Cogent and submit their own production report to T&E for rating. As a result, Co-Arg users generated more reports per challenge round than any of the other evaluated team-based system, as can be seen in Table 33. The systems where users worked individually – control/individual (Conclude) and TRACE - submitted one report for each user, resulting in more reports overall. Users could submit their reports more than one time for the same problem (successive versions of the report while they worked on it), but only the last submission was considered for rating.

		ctrl_team	bard	coarg	swarm	*Total*
Practice	pulse	5	.	.	.	*5*
	rakona	4	.	.	.	*4*
	cellnet	.	9	.	6	*15*
	kalukistan_bomb	.	.	.	8	*8*
	manada_sam_sale	.	.	25	.	*25*
	smoking_and_cancer	.	12	.	.	*12*
	salazar	.	.	22	.	*22*
	Total	*9*	*21*	*47*	*14*	*91*
	Average		*11.57*	*6.62*	*23.14*	
Round 1	black_site_surveillance	1	4	8	2	*15*
	missile-test	1	2	7	1	*11*
	lifehacker	1	1	7	2	*11*
	arthur_allen	1	1	10	3	*15*
	Total	*4*	*8*	*32*	*8*	*52*
	Average	*8*	*13.75*	*3.03*	*18.13*	
Round 2	cyberattack	1	3	5	1	*10*
	zest	1	1	4	1	*7*
	which_lovell	1	1	5	1	*8*
	who-is-the-spy	1	1	8	2	*12*
	Total	*4*	*6*	*22*	*5*	*37*
	Average	*4.5*	*8.17*	*2.27*	*16.4*	
Round 3	frogs	1	1	4	1	*7*
	without_a_trace	1	1	4	1	*7*
	dark_web	1	2	3	1	*7*
	prison_break	1	1	4	1	*7*
	Total	*4*	*5*	*15*	*4*	*28*
	Average	*4.5*	*6.8*	*2.33*	*16*	

Table 32: Team Commitment (from data provided by T&E on August 27, 2018).

		bard	coarg	ctrl_ind	ctrl_team	swarm	trace	*Total*
Round 1	arthur_allen	1	7	24	1	2	32	*67*
	black_site_surveillance	4	2	9	1	1	14	*31*
	lifehacker	1	2	16	1	1	25	*46*
	missile_test	2	5	22	0	1	19	*49*
	Total	*8*	*16*	*71*	*3*	*5*	*90*	*193*
Round 2	cyberattack	3	2	12	0	1	5	*23*
	zest	1	2	8	1	1	16	*29*
	which_lovell	1	4	15	1	2	15	*38*
	who-is-the-spy	1	1	15	1	0	6	*24*
	Total	*6*	*9*	*50*	*3*	*4*	*42*	*114*
Round 3	dark_web	1	4	10	1	1	15	*32*
	frogs	1	3	3	1	1	8	*17*
	prison_break	1	2	8	1	1	12	*25*
	without_a_trace	1	2	8	2	0	7	*20*
	Total	*4*	*11*	*29*	*5*	*3*	*42*	*94*

Table 33: Ratable Products (from data provided by T&E on August 27 and 29, 2018).

8.2. GMU Problems in the T&E Experiment

The T&E experiment used two of the problems proposed by GMU, “Fillistan Conducts Ballistic Missile Tests” in Round 1 (see Appendix 9.7), and “Who is the Spy?” in Round 2 (see Appendix 9.9). In our view, these are the most difficult of all the 20 problems selected to be used in the T&E evaluation.

“Fillistan Conducts Ballistic Missile Tests” meets all the 8 key elements of high-quality analytic reasoning identified by IARPA and T&E, including “generation of unique analytic insights.” The answer to the intelligence question requires the identification of a hypothesis that is not obvious and that cannot be identified without detailed analysis of the more obvious possible hypotheses, each of which is unlikely for a different reason. Problem also requires extensive analysis of source credibility as some sources are totally unreliable and are being used in a denial-and-deception (D&D) effort.

“Who is the Spy?” also all the 8 key elements of high-quality analytic reasoning, including “generation of unique analytic insights.” The answer to the intelligence question is not possible without modification of each of the hypotheses to resolve what could be an analytic dead end. The answer to the question also requires the development of an analytic framework prior to evaluating individual pieces of information “to identify the most likely and persuasive objections to key judgments, hypotheses, conclusions, and assumptions” as outlined in the criterion “identification, quality assessment, and refutation of potential objections.”

The following sections discusses some of the solutions provided by the participants in the T&E experiment.

8.3. Discussion of the Solutions of the “Fillistan” Problem

The T&E experiment resulted in five solutions with Co-Arg of the “Fillistan Conducts Ballistic Missile Tests” problem presented in Appendix 9.7. One of these five solutions (dolphinfin_Fillistan_missile) is presented in Appendix 9.8. *This solution is was one of the best that we have seen for this difficult problem, including in our internal testing.* This solution was the only one to identify and assess an unknown missile as a possible hypothesis. All of the other solutions limited the analysis to an evaluation of the known missiles that were identified in the problem’s available information. The analysis in this exemplary solution linked several disparate reports involving institute directors and institute locations to conclude that a new, unidentified missile was being tested. In addition, this particular solution was also one of the few in all of our testing that doubted the credibility of the human sources after a certain period.

The argumentations in the other four reports on the missile-test problem were well done in many areas, but because these reports failed to consider an unknown missile, the accuracy of their solutions was low. All of the reports demonstrated a serious effort to solve this difficult problem in a limited time using Co-Arg to help guide their argumentation.

All four solutions that included Cogent diagrams in the missile problem—one solution did not include any Cogent diagrams in the submitted report—showed a capability to use Cogent correctly, *demonstrating that Cogent can be learned and used relatively easily.* Users in the T&E experiment clearly understood and properly assessed the credibility and relevance factors used in Co-Arg, and demonstrated the ability to correctly construct both “and” and “or” arguments, as well as arguments with favoring and disfavoring evidence.

The diagram in Figure 47 is from the “aerojdkatz_Fillistan_1” solution. The analysis is an “or” argument and correctly diagrams that if <u>either</u> the range or burn time of the Victory missile is inconsistent with the tested missile, this information becomes the basis for an argument against the hypothesis that the tested missile was a Victory missile.

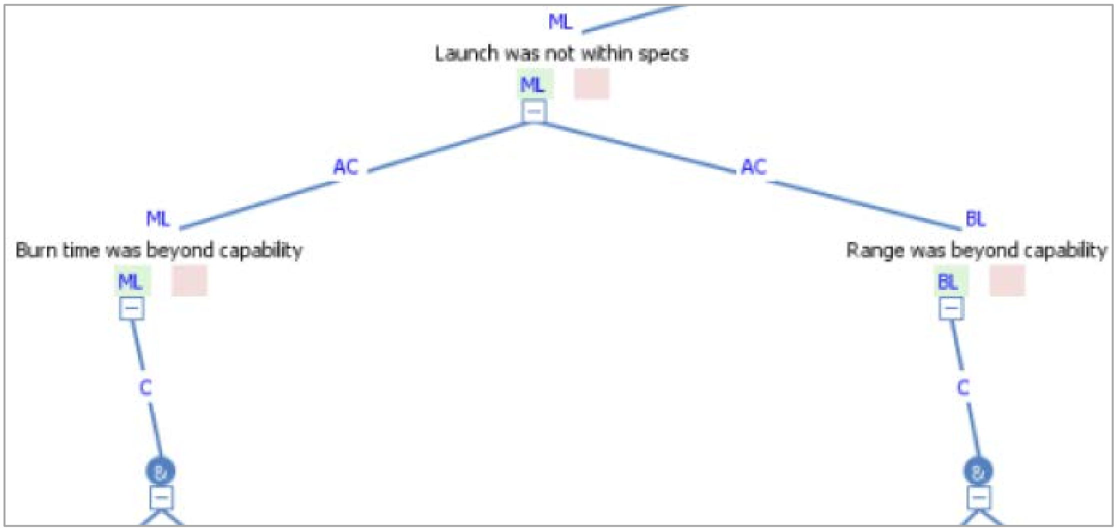
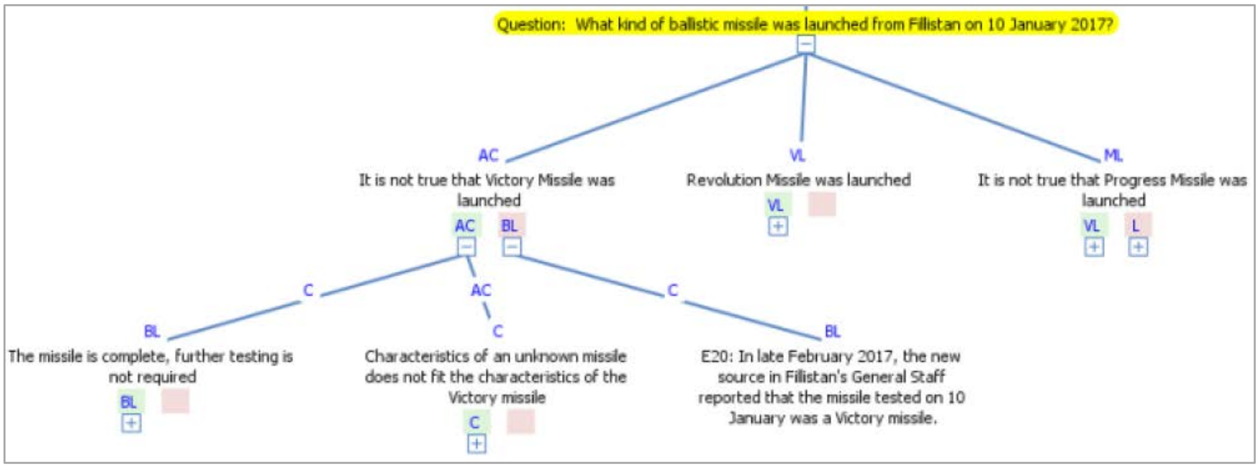

ML
Launch was not within specs
ML
AC
AC
ML
Burn time was beyond capability
ML
BL
Range was beyond capability
BL
C
C

Figure 47: Multiple favoring arguments in the “aerojdkatz_Fillistan_1” solution.

In the diagram from Figure 48, which was taken from the "llhh13_Fillistan_conducts" solution, the top-level analysis of the likelihood that the missile was not a Victory missile includes favoring arguments related to the missile's specifications and the lack of a need to test a reliable, already developed missile, as well as disfavoring information from a source who reported that it was a Victory missile. The analysis in this report correctly assessed the disfavoring information as outweighing the information from the human source (who had been compromised). The diagram shows that the user for this solution was at least as focused on disconfirming evidence as confirming evidence. The argumentation in most T&E user solutions shows that users were constructing arguments that evaluated all the evidence, as opposed to selectively choosing confirming evidence to support a preferred hypothesis. Note also that the user correctly assessed the relevance of the information from the human source as "certain" (C). If this information was true, the hypothesis that it was not true that it was the Victory had to be false.



Question: What kind of ballistic missile was launched from Fillistan on 10 January 2017?

It is not true that Victory Missile was launched

Revolution Missile was launched

It is not true that Progress Missile was launched

The missile is complete, further testing is not required

Characteristics of an unknown missile does not fit the characteristics of the Victory missile

E20: In late February 2017, the new source in Fillistan's General Staff reported that the missile tested on 10 January was a Victory missile.

Figure 48: Favoring and disfavoring arguments in the "llhh13_Fillistan_conducts" solution.

In the diagram from Figure 49, taken from the "dolphinsfin_Fillistan_missile" solution, the analysis explains a key sub-judgment that was necessary to support the hypothesis that an unknown missile was being tested. The analysis used information from three different reports to construct a multi-tiered "and" argument.

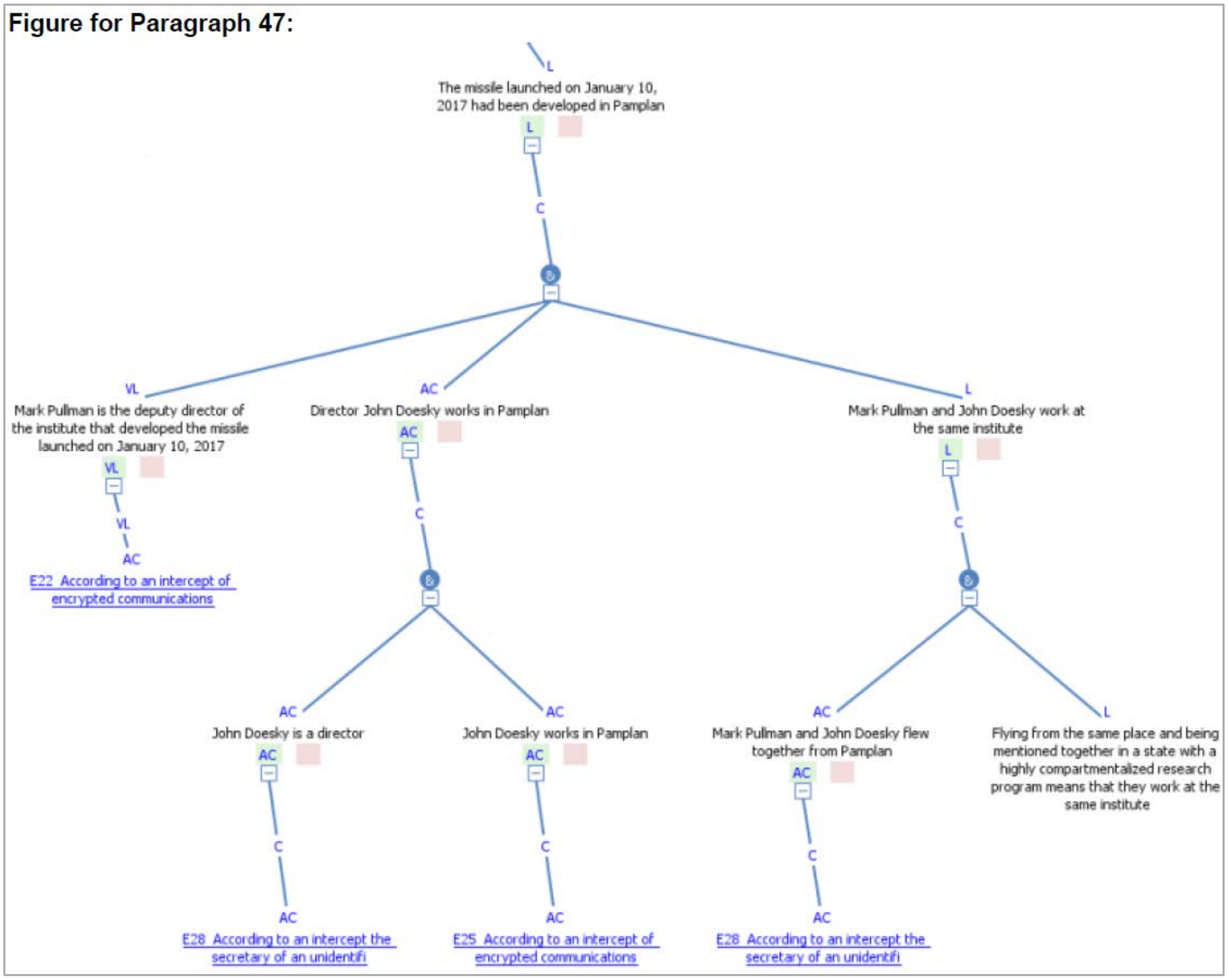


Figure for Paragraph 47:
L
The missile launched on January 10, 2017 had been developed in Pamplan
L
C
&
VL
Mark Pullman is the deputy director of the institute that developed the missile launched on January 10, 2017
VL
VL
AC
E22 According to an intercept of encrypted communications
AC
Director John Doesky works in Pamplan
AC
C
&
L
Mark Pullman and John Doesky work at the same institute
L
C
&
AC
John Doesky is a director
AC
C
AC
E28 According to an intercept the secretary of an unidentifi
AC
John Doesky works in Pamplan
AC
C
AC
E25 According to an intercept of encrypted communications
AC
Mark Pullman and John Doesky flew together from Pamplan
AC
C
AC
E28 According to an intercept the secretary of an unidentifi
L
Flying from the same place and being mentioned together in a state with a highly compartmentalized research program means that they work at the same institute

Figure 49: Multi-tired "and" argumentation in the "dolphinsfin_Fillistan_missile" solution.

8.4. A Solution of the "Fillistan" Problem in the T&E Experiment, and its Rating

Appendix 9.8 contains the solution "dolphinfin_Fillistan_missile" provided by a participant in the T&E evaluation for this problem.

The following table contains the quality of reasoning evaluation grid we developed for this problem, and shows the points received by the participant's solution for each dimension of quality of reasoning. Notice that this solution received 31.5 points out of 39 (8.08 on a 10 point scale), probably the best seen on this very difficult problem.

Fillistan Conducts Ballistic Missile Tests	
Criterion	**Points**
	Total: 39 – (Awarded 31.5)
Hypothesis generation and accuracy of solution	Subtotal 9 – **(6.5)**
Key Judgments: • Missile tested on 10 January was the Victory • Missile tested on 10 January was the Revolution • Missile tested on 10 January was the Progress • Missile tested on 10 January was an unidentified missile in development	4 (1 each) – **(4)**
Accuracy of solution: • Unknown missile very likely was tested on 10 January Note: dolphinsfin_Fillistan_missile, which judged that a new solid-fueled missile **likely (55-70%)** was tested, was very close to receiving points for accuracy	1 point for very likely (80-95%); 0.5 for more than likely (70-80%) or almost certainly (95-99%) – **(0)**
• There is almost no chance that the Victory missile was tested	1 point for very unlikely (5-20%); 0.5 for almost no chance (1-5%) or more than unlikely (20-30%); 0.25 points for lacking support (0-50%) -- **(1)**
• The Revolution missile very unlikely (5-20%) was tested	1 point for very unlikely (5-20%); 0.5 for almost no chance (1-5%) or more than unlikely (20-30%); 0.25 points for lacking support (0-50%) – **(0)**
• The Progress missile more than unlikely (20-30%) was tested	1 point for more than unlikely (20-30%); 0.5 very unlikely (5-20%) or unlikely (30-45%); 0.25 points for lacking support (0-50%) – **(.5)**
Numerical probability is consistent with qualitative response	1 – **(1)** (0.25 points for each hypothesis)
Argument structure and reasoning	Subtotal: 14 — **(10)**
Reasoning behind unknown missile • (For) The designer working on the missile tested in January and the one that exploded in March was working at an unidentified institute in Pamplan and not the institutes associated with the Victory, Revolution, or Progress missiles • (For) Fillistan has a very compartmented missile design and development process so there is no overlap between institutes • (For) Fillistan allocated a large amount of funding for development of solid-fueled missiles, more than enough for just one such missile • (Against) A source says Fillistan has no new missiles in development	4 (1 each) – **(3)**
Reasoning behind Progress missile: • (For) Progress is under development and is suitable for testing at Matana missile and test development facility • (Against) Development as of late December was not expected to be completed until March so would not have been ready for testing in January	3 (1 each) – **(2)**

• (Against) Development as of 8 January making progress but official did not note development was completed	
Reasoning behind Revolution: • (For) The capabilities of the tested missile (range and engine operation time) were consistent with the Revolution • (For) Testing at Matana missile development and test facility consistent with Revolution missile undergoing modifications • (Against) The missile tested on 10 January was the same missile that exploded on 20 March and the Revolution is a reliable missile that would not have failed • (Against) Institute developing missile tested on 10 January was developing its first missile but the Revolution is already an operational missile	4 (1 each) – **(2)**
Reasoning behind Victory missile • (For) A source says it was the Victory missile • (Against) The capabilities of the missile (range and engine operation time) tested on 10 January are different than those of the Victory • (Against) Testing at Matana missile development and test facility inconsistent with operational missile that has not had any performance issues **and/or** military very happy with performance of missile	3 (1 each) – **(3)**
Assessment of credibility of evidence and identification of sources	Subtotal: 12 – **(12)**
Considers information from all non-human sources credible	2 – **(2)**
Considers information from human sources prior to December 2016 as credible	2 – **(2)**
Identifies ALL human sourcing after March as not credible as sources likely were compromised by late 2016 or early 2017	2 – **(2)**
Considers reporting from source that no new solid-fueled missiles were in development and/or that Hall received the "Hero" award for work on Revolution and/or that as Victory was tested on 10 January not credible	2 – **(2)**
Sourcing:	
• Provides extensive source references	4 – (4)
• Provides few source references:	2
• Provides hardly any source references	0
Identification of key missing information and assumptions	Subtotal: 4 – **(3)**
• No information available on a second solid-fuel missile; assumption: if the missile tested on 10 January was not the Victory, Revolution, or Progress, then it was a new missile not previously identified	3 – **(3)**
• It was not known whether development of Progress was accelerated enough to be ready for testing on 10 January; assumption: if the missile was going to be tested in two days, on 10 January, the engineer in intercept would have said so on 8 January instead of saying work on the tank was going faster than expected	1 – **(0)**

8.5. A Solution of the "Spy" Problem in the T&E Experiment, and its Rating

The T&E experiment resulted in four solutions with Co-Arg of the "Who is the Spy?" problem presented in Appendix 153.

One of those four solutions, "aaron83m_who_is_the_spy," was quite well done—scoring 29 out of 47 points (6.17 on a 10 points scale) —and arguably might have done much better in our evaluation had some ambiguity in this problem that we inadvertently created been eliminated.

- The author clearly understood that the data showed that the three military officers with access had strong alibis for at least one of the days. How did the author handle this: he noticed that Harry did not have an alibi on two of the days and then took the information "Harry was in office at noon" and document passed at "around noon" and interpreted "around noon" as allowing enough wiggle room to deliver the document. If the language in the problem was more definitive, such as Harry was in the office "all day" as it was for the other two days information was passed, the author might have started looking for another explanation, such as having an accomplice, a key finding necessary for an accurate solution.
- The author of this solution also identified Harry's lack of a personal relationship—Harry was a complete loner in this problem—with anyone in Arboria as an argument for Harry being capable of betraying Arboria's trust. This is a good argument but was not considered in the school solution, when it arguably could have. We believe this demonstrates the ability of Cogent, which requires the development of a complete argument, to improve and develop argumentation at specific levels in the argument.

In the analysis related to whether the custodian (identified in this problem as a possible suspect) was passing the documents, the report "aaron83m_who_is_the_spy" weighed and correctly diagrammed the favoring evidence (needs the money) against the disfavoring evidence (problematic access, a clean record for decades with no criminal conduct, and a likely lack of familiarity with details in the note that accompanied the documents). The user also provided a defensible assessment for the relevance of these sub-reasons. For example, the relevance of the sub-reason on no recent criminal record was only assessed as "likely"—defensible because it allows for the possibility that the custodian reverting to his criminal tendencies.

The following table contains the quality of reasoning evaluation grid we developed for this problem, and shows the points received by "aaron83m_who_is_the_spy" for each dimension of quality of reasoning.

Who's the Spy?	
Criterion	**Points**
	Total 47 – (Awarded 29)
Hypothesis generation and accuracy of solution	Subtotal 9 – (6)
Addresses each of the following hypotheses: • Tom removed the information • Harry removed the information • Dick removed the information • Paul removed the information	4 (1 each) – (4)
Solution: • We assess it is barely likely that Tom removed the classified information about Plumistan that was passed to Razmania's Embassy	1 point for barely likely (50-55%); 0.5 for likely (55-70%) or barely unlikely (45-50%) – (0)
• We assess it is unlikely that Harry removed the classified information about Plumistan that was passed to Razmania's Embassy	1 point for unlikely (30-45%); 0.5 for barely unlikely (45-50%) or more than unlikely (20-30%) – (0)
• We assess it is very unlikely that Dick removed the classified information about Plumistan that was passed to Razmania's Embassy (Note: Concluded there was no chance that Dick was the person—very close to receiving .5 points.)	1 point for very unlikely (5-20%); 0.5 for more than unlikely (20-30%) or almost no chance (1-5%)— (0)
• We assess it is very unlikely that Paul removed the classified information about Plumistan that was passed to Razmania's Embassy	1 point for very unlikely (5-20%); 0.5 for more than unlikely (20-30%) or almost no chance (1-5%) — (1)
• Numerical probability is consistent with qualitative response	1 — (1) (0.25 points for each hypothesis)
Argument structure and reasoning	**Subtotal 29 – (17)**
Reasoning in "Tom" argumentation: • (Favoring as having motive) Tom was very angry with the Navy for not being promoted • (Favoring as having motive) Tom was under pressure from his wife to earn more money to meet her desired living standards • (Favoring as having motive) Tom's credit cards very likely were maxed out • (Favoring as being knowledgeable) Tom had the background and knowledge that the individual passing the information displayed. He would have been familiar with the mathematician Euclid and known what information would have been useful for Razmania and that Mailandia had broken Razmania's codes • (Favoring as predisposed to breaking rules) Tom recently disregarded procedures for foreign travel • (Disfavoring as predisposed to breaking rules) Tom previously had a track record of being a straight arrow • (Favoring as having opportunity) Only Tom had an obvious potential accomplice, his wife Karen who was angrier than Tom was when he was not promoted	8 (1 each) – **(5)**

• (Disfavoring as having opportunity) Tom was at work in the CCC during at least one of the times the information was passed but could have used an accomplice	
Reasoning in “Harry” argumentation: • (Favoring as having motive) Harry had a love of Razmania and criticized Mailandian policy and also had regular contacts with Mailandians • (Favoring as predisposed to breaking rules) Harry failed to report one contact but he is known to be absent-minded and was working on a highly demanding project at the time • (Disfavoring as predisposed to breaking rules) Harry has reported all his other contacts and travel • (Favoring as being knowledgeable) Harry had the background and knowledge that the individual passing the information displayed. He would have been familiar with the mathematician Euclid and known what information would have been useful for Razmania and that Mailandia had broken Razmania’s codes • (Disfavoring as having opportunity) Harry was in the DCA on 9 December when information was passed • (Disfavoring as having opportunity) Harry was known to be a loner with no friends but his cat, so he did not have an obvious accomplice to pass information while he was at work on 9 December	6 (1 each) – **(3)**
Reasoning in “Dick” argumentation: • (Favoring as having motive) Dick was also dealing with financial burdens • (Favoring as having motive) Dick had also been passed up for promotion • (Disfavoring as having motive) Dick had recently quit smoking and drinking and appeared upbeat about his future promotion prospects • (Favoring as being predisposed to breaking rules) Dick was willing to break the rules by having an adulterous affair • (Favoring as being knowledgeable) Dick had the background and knowledge that the individual passing the information displayed. He would have been familiar with the mathematician Euclid and known what information would have been useful for Razmania and that Mailandia had broken Razmania’s codes • (Disfavoring as having opportunity) Dick was in the CCC at least during one of the times the information was passed • (Disfavoring as having opportunity) Dick was in the process of divorcing his wife and possibly breaking up with his girlfriend and his best and only friend was out of the country at the time the information was passed	7 (1 each) – **(5)**
Reasoning in “Paul” argumentation: • (Favoring as having motive) Paul was in need of money for a kidney transplant for his daughter	8 (1 each) – **(4)**

• (Favoring as having opportunity) Paul was the only one of the four who was not in the DCA building during at least one of the times when information was passed • (Disfavoring as having opportunity) Paul was escorted and monitored closely while in the vault • (Favoring as being predisposed to breaking rules) Paul had a criminal past • (Disfavoring as being predisposed to breaking rules) Paul has not been in trouble with the law since getting out of prison • (Disfavoring as being knowledgeable) Paul did not have clearances and would not have known that Mailandia had broken Razmania's codes • (Disfavoring as being knowledgeable) Paul with only a 10th grade education and who spent his free time watching reality shows would not have known who Euclid was • (Disfavoring as being knowledgeable) Paul would not have known about Razmania's security interests	
Assessment of quality and credibility of evidence	**Subtotal 6 – (6)**
Considers all of the evidence as credible—no credibility issues noted in the production report	2 – **(2)**
Sourcing:	
• Provides extensive source references	4 – **(4)**
• Provides few source references	2
• Provides hardly any source references	0
Identification of key missing information and assumptions	**Subtotal 3 – (0)**
Key missing information: three of suspects needed an accomplice	3

8.6. Conclusions

We believe that these solutions represent a proof of concept that Co-Arg can be used to solve complex problems with a limited amount of training. The dexterity that participants demonstrated in using Co-Arg after only two hours of required training was impressive.

The results also suggest that our assumption that students would be motivated by the class grade to diligently learn and use Co-Arg may be wrong. We provided five practice problems to these students, and all were required. However, as discussed in Section 4.5, on average, these participants fully completed only 2.4 practice problems. We can only speculate on this lack of commitment, but one possibility is that many students judged that their grade for the class would still be acceptable (no lower than a "B") regardless of how they performed in the experiment.

The participants in the T&E experiment appear to have been quite motivated, and obtained better results than the students who had much more training. Again, we can only speculate on what appears to be a greater commitment to learn and use Co-Arg, but one possibility is that participants in the T&E experiment chose Co-Arg themselves, in part because of the introductory video that highlighted Co-Arg's analytic advantages. The relatively large number of participants that opted for Co-Arg over other systems in the T&E experiment tends to support this conclusion.

Intelligence analysts are motivated professionals and if they perceive that a tool is helping them perform their jobs, they are likely to use it. Many analysts' may have a bias against training partly because they perceive that too much of the training they take is not overly helpful and there is not enough of a "return" for their investment in time. If this is true and if analysts perceive that Co-Arg provides significant analytic advantages—which we believe is likely—the premise that analysts would embrace a tool that has only very minimum training requirements may not be accurate.

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under the CREATE contract number 2017-16112300009, and by George Mason University. The IARPA CREATE program was managed by Steven Rieber, supported by Ruthana Gordon. Richard Lempert provided helpful advice on improving the Co-Arg system and the associated user training. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, 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.

Chirag Utamsingh, Terri Galanti, Joan McIntyre, Vivian Motti, Hui Zheng and Manpria Dua contributed to this research.

Mark Clark, Cindy Ayers, Jody Neathery-Castro, Lana Obradovic, Eric Gable, and Tom Winston contributed to the evaluation of Co-Arg.

References

Bangor, A., Kortum, P., and Miller, J. (2009). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale, Journal of Usability Studies, 4(3), pp.114-123.

Bozzon, A., Brambilla, M., Ceri, S., Silvestri, M., and Vesci, G. (2013). Choosing the right crowd: Expert finding in social networks. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, Genoa, Italy, March 18-22, pp. 637-648.

Chen, C., White, L., Kowalewski, T., Aggarwal, R., Lintott, C., Comstock, B., Kuksenok, K., Aragon, C., Holst, D., and Lendvay, T. (2013). Crowd-Sourced Assessment of Technical Skills: a novel method to evaluate surgical performance. *Journal of Surgery Research*. 2014 March. 187(1), pp.65-71. Epub 2013 Oct 10.

Cohen, L. J. (1977), *The Probable and the Provable,* Clarendon Press, Oxford.

Cohen, L. J. (1989). *An Introduction to the Philosophy of Induction and Probability*, Clarendon Press, Oxford.

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

Heuer, R. J. (2008). Computer-Aided Analysis of Competing Hypotheses, in George R. Z., Bruce J. B., eds., *Analyzing Intelligence: Origins, Obstacles, and Innovations*, Georgetown University Press, Washington, DC.

Heuer, R. J., and Pherson R. H. (2011). *Structured Analytic Techniques for Intelligence Analysis,* CQ Press, Washington, DC.

Howe, J. (2006). The Rise of Crowdsourcing. *Wired*. Available online: https://www.wired.com/2006/06/crowds/ (accessed September 20, 2018).

Hugin (2016). http://www.hugin.com/ (accessed May 3, 2016)

IARPA (2016). Crowdsourcing Evidence, Argumentation, Thinking and Evaluation, BAA, https://www.iarpa.gov/index.php/research-programs/create/create-baa

ICD 203. (2007). Intelligence Community Directive 203, ODNI.

Lakhani, Karim R., David A. Garvin, and Lonstein E. (2010). TopCoder (A): Developing Software through Crowdsourcing. *Harvard Business School Case 610-032*, January 2010. (Revised May 2012.)

Maarry, K.E., Balke, W.T., Cho, H., Hwang, S., and Baba, Y. (2014). Skill Ontology-Based Model for Quality Assurance in Crowdsourcing. In Han, W.S., Lee, M.L., Muliantara, A., Saniaya, N.A., Thalheim, B., and Zhou, S., *Database Systems for Advanced Applications*, Series Lecture Notes in Computer Science, ISBN: 978-3-662-43983-8, Vol 8505, Spinger, 2014, pp. 376-387.

Mavridis, P., Gross-Amblard, D., and Miklós, Z. (2016). Using Hierarchical Skills for Optimized Task Assignment in Knowledge-Intensive Crowdsourcing. *Proceedings of the 25th International Conference on World Wide Web*, Montreal, Canada, April 11-15, pp. 843-853.

Negoita, C. V., and Ralescu, D. A. (1975). *Applications of Fuzzy Sets to Systems Analysis*, Wiley, New York.

Netica (2016). http://www.norsys.com/ (accessed May 3, 2016)

Nilsson, N.J. (1971). *Problem Solving Methods in Artificial Intelligence.* NY: McGraw-Hill.

Pope S., and Josang A. (2005). Analysis of Competing Hypotheses using Subjective Logic (ACH-SL), Queensland University, Brisbane, Australia, ADA463908.

Rationale (2016). http://www.reasoninglab.com/learn/ (accessed May 3, 2016)

Reagle, J.M. Jr. (2010) *Good Faith Collaboration, The Culture of Wikipedia.* MIT Press. ISBN: 9780262014472. Available online: https://mitpress.mit.edu/books/good-faith-collaboration (accessed September 20, 2018).

Russell S.J., and Norvig P. (2010). Artificial Intelligence: A Modern Approach, Prentice-Hall.

Safire, W. (2009). On Language. *New York Times Magazine*. Available online: https://www.nytimes.com/2009/02/08/magazine/08wwln-safire-t.html (accessed September 20, 2018).

SEP (2016). The Stanford Encyclopedia of Philosophy. Available online: http://plato.stanford.edu/ (accessed September 20, 2018).

Schum, D. (1987). *Evidence and Inference for the Intelligence Analyst* (Two Volumes). Lanham, MD: University Press of America.

Schum, D.A. (2001). *The Evidential Foundations of Probabilistic Reasoning*, Northwestern University Press

Simonite, T. (2013). The Decline of Wikipedia. MIT Technology Review. October 22. Available online: https://www.technologyreview.com/s/520446/the-decline-of-wikipedia/ (accessed September 20, 2018).

Shafer, G. (1976). *A Mathematical Theory of Evidence,* Princeton University Press, Princeton, NJ.

Sonnad, N. (2015). This free online encyclopedia has achieved what Wikipedia can only dream of. *Quartz*, September 21. http://qz.com/480741/this-free-online-encyclopedia-has-achieved-what-wikipedia-can-only-dream-of (accessed September 20, 2018).

Tecuci, G. (1988). *DISCIPLE: A Theory, Methodology and System for Learning Expert Knowledge.* Thèse de Docteur en Science, University of Paris-South, France.

Tecuci, G. (1998). *Building Intelligent Agents: An Apprenticeship Multistrategy Learning Theory*, Methodology, Tool and Case Studies. London, England: Academic Press.

Tecuci, G. (2012). Artificial Intelligence, *Wiley Interdisciplinary Reviews: Computational Statistics*, 4: 168-180, doi: 10.1002/wics.200.

Tecuci, G., Boicu, M., Ayers, C., and Cammons D. (2005). Personal Cognitive Assistants for Military Intelligence Analysis: Mixed-Initiative Learning, Tutoring, and Problem Solving. In *Proc. 1st Int. Conf. on Intelligence Analysis*, McLean, VA.

Tecuci, G., Boicu, M., and Cox, M.T. (2007a). Seven Aspects of Mixed-Initiative Reasoning: An Introduction to the Special Issue on Mixed-Initiative Assistants, AI Magazine, 28 (2), pp. 11-18, Summer.

Tecuci, G., Boicu, M., Marcu, D., Boicu, C., Barbulescu, M., Ayers, C., and Cammons, D. (2007b). Cognitive Assistants for Analysts, *Journal of Intelligence Community Research and Development* (JICRD). Also in Auger J. and Wimbish W. eds. *Proteus Futures Digest*, 303-329, Joint publication of National Intelligence Univ., Office of the Director of National Intelligence, and US Army War College Center for Strategic Leadership.

Tecuci, G., Boicu, M., Marcu, D., Boicu, C., and Barbulescu, M. (2008). Disciple-LTA: Learning, Tutoring and Analytic Assistance, *Journal of Intelligence Community Research and Development.*

Tecuci, G., Marcu, D., Boicu, M., Schum, D. A., and Russell K. (2011). "Computational Theory and Cognitive Assistant for Intelligence Analysis," *Proceedings of the Sixth International Conference on Semantic Technologies for Intelligence, Defense, and Security – STIDS*, pp. 68-75, Fairfax, VA, 16-18 November.

Tecuci, G., Schum, D.A., Marcu, D., Boicu, M., Computational Approach and Cognitive Assistant for Evidence-Based Reasoning in Intelligence Analysis, *International Journal of Intelligent Defence Support Systems,* Vol. 5, No. 2, pp. 146-172, 2014.

Tecuci, G., Marcu, D., Boicu, M., and Schum, D.A. (2015). COGENT: Cognitive Agent for Cogent Analysis, in *Proceedings of the 2015 AAAI Fall Symposium "Cognitive Assistance in Government and Public Sector Applications,"* 58-65, Arlington, VA, Technical Report FS-15-02, AAAI Press, Palo Alto, CA.

Tecuci G., Marcu D., Boicu M., and Schum D.A. (2016a). *Knowledge Engineering: Building Cognitive Assistants for Evidence-based Reasoning*, Cambridge University Press.

Tecuci, G., Schum, D. A., Marcu, D., and Boicu, M. (2016b). *Intelligence Analysis as Discovery of Evidence, Hypotheses, and Arguments: Connecting the Dots,* Cambridge University Press.

Tecuci, G., Kaiser, L., Marcu, D., Uttamsingh, C., and Boicu, M. (2018). Evidence-based Reasoning in Intelligence Analysis: Structured Methodology and System, Special Issue on Evidence-based Reasoning and Applications, *Computing in Science and Engineering,* Vol. 20, Issue 6, pp. 9-21.

Toulmin S. E. (1963). *The Uses of Argument*, Cambridge University Press.

Valtorta M., Dang J., Goradia H., Huang J., and Huhns M. (2005). Extending Heuer's Analysis of Competing Hypotheses Method to Support Complex Decision Analysis. *Proceedings of the International Conference on Intelligence Analysis (IA-05)*, Washington, D.C., May 2-4.

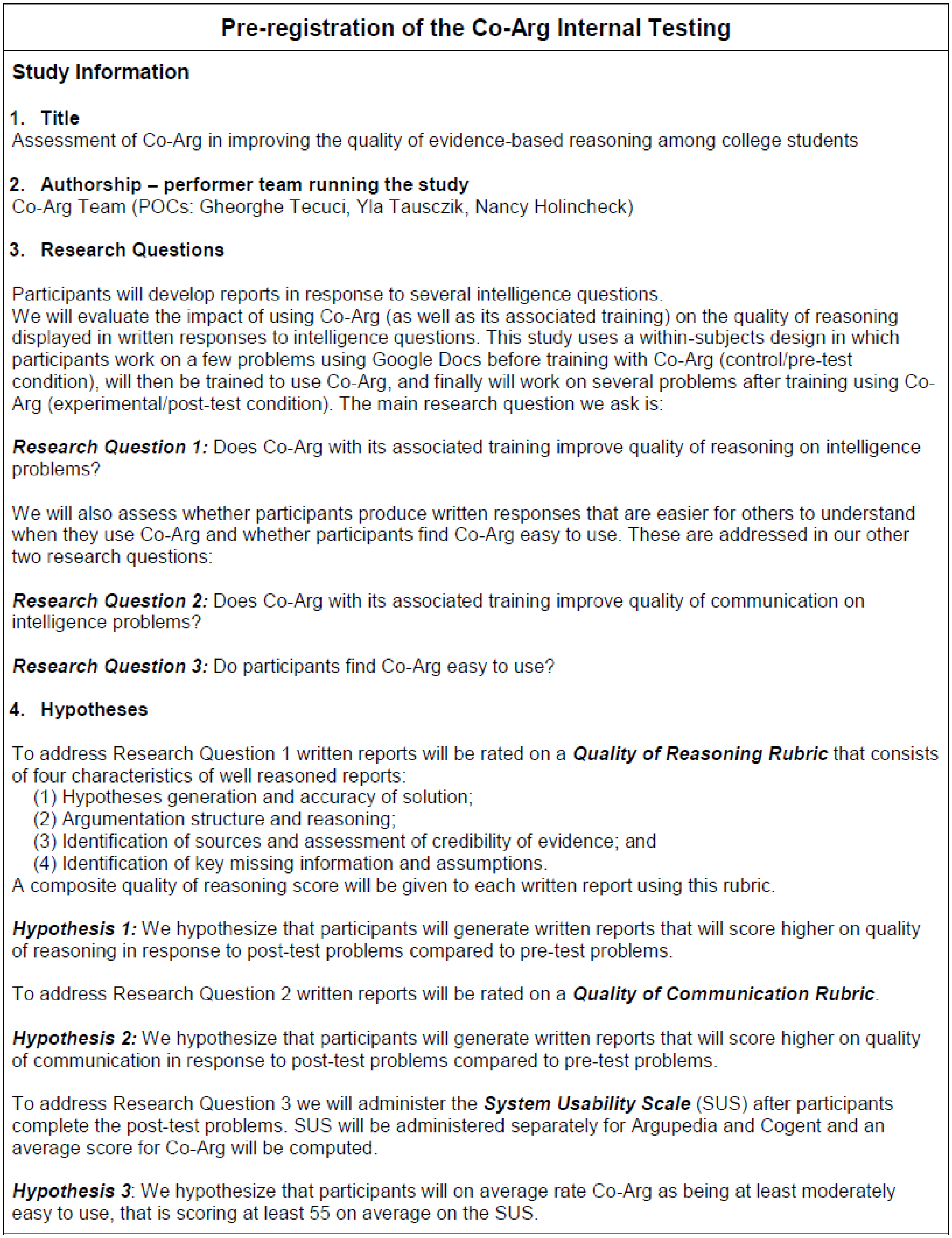
van Gelder T.J. (2007). The Rationale for Rationale, *Law, Probability and Risk*, 6, pp.23-42.

Wheaton K.J., and Chido D.E. (2006). Structured Analysis of Competing Hypotheses: Improving a Tested Intelligence Methodology, *Competitive Intelligence Magazine*, 9 (6), pp.12–15.

Zadeh, L. (1983). The Role of Fuzzy Logic in the Management of Uncertainty in Expert Systems, *Fuzzy Sets and Systems*, Vol. 11, pp. 199-227.

9. Appendixes

9.1. Pre-registration of the Co-Arg Internal Testing



Pre-registration of the Co-Arg Internal Testing

Study Information

1. **Title**

Assessment of Co-Arg in improving the quality of evidence-based reasoning among college students

2. **Authorship – performer team running the study**

Co-Arg Team (POCs: Gheorghe Tecuci, Yla Tausczik, Nancy Holincheck)

3. **Research Questions**

Participants will develop reports in response to several intelligence questions.
We will evaluate the impact of using Co-Arg (as well as its associated training) on the quality of reasoning displayed in written responses to intelligence questions. This study uses a within-subjects design in which participants work on a few problems using Google Docs before training with Co-Arg (control/pre-test condition), will then be trained to use Co-Arg, and finally will work on several problems after training using Co-Arg (experimental/post-test condition). The main research question we ask is:

Research Question 1: Does Co-Arg with its associated training improve quality of reasoning on intelligence problems?

We will also assess whether participants produce written responses that are easier for others to understand when they use Co-Arg and whether participants find Co-Arg easy to use. These are addressed in our other two research questions:

Research Question 2: Does Co-Arg with its associated training improve quality of communication on intelligence problems?

Research Question 3: Do participants find Co-Arg easy to use?

4. **Hypotheses**

To address Research Question 1 written reports will be rated on a ***Quality of Reasoning Rubric*** that consists of four characteristics of well reasoned reports:
(1) Hypotheses generation and accuracy of solution;
(2) Argumentation structure and reasoning;
(3) Identification of sources and assessment of credibility of evidence; and
(4) Identification of key missing information and assumptions.
A composite quality of reasoning score will be given to each written report using this rubric.

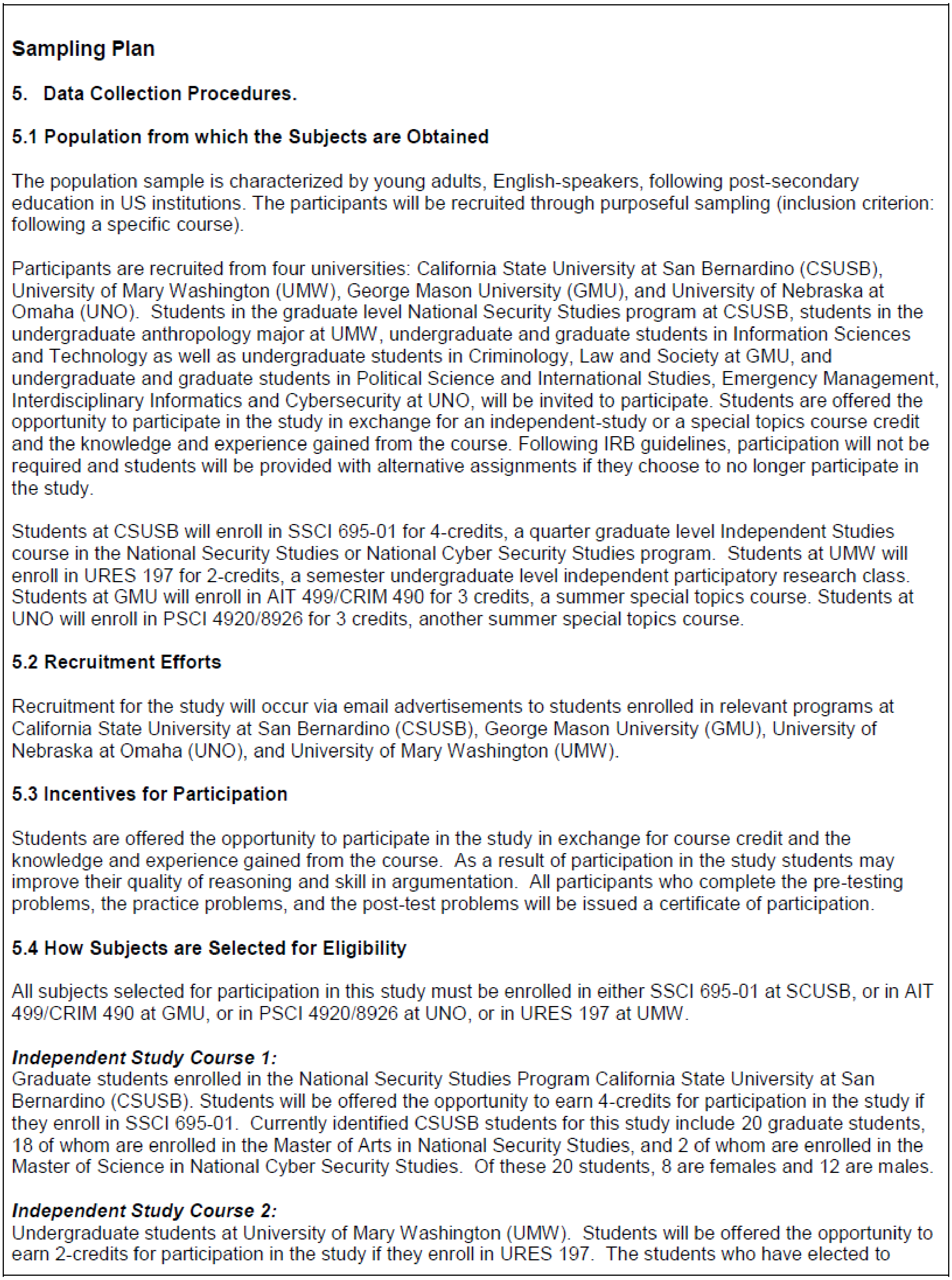
Hypothesis 1: We hypothesize that participants will generate written reports that will score higher on quality of reasoning in response to post-test problems compared to pre-test problems.

To address Research Question 2 written reports will be rated on a ***Quality of Communication Rubric***.

Hypothesis 2: We hypothesize that participants will generate written reports that will score higher on quality of communication in response to post-test problems compared to pre-test problems.

To address Research Question 3 we will administer the ***System Usability Scale*** (SUS) after participants complete the post-test problems. SUS will be administered separately for Argupedia and Cogent and an average score for Co-Arg will be computed.

Hypothesis 3: We hypothesize that participants will on average rate Co-Arg as being at least moderately easy to use, that is scoring at least 55 on average on the SUS.



Sampling Plan

5. Data Collection Procedures.

5.1 Population from which the Subjects are Obtained

The population sample is characterized by young adults, English-speakers, following post-secondary education in US institutions. The participants will be recruited through purposeful sampling (inclusion criterion: following a specific course).

Participants are recruited from four universities: California State University at San Bernardino (CSUSB), University of Mary Washington (UMW), George Mason University (GMU), and University of Nebraska at Omaha (UNO). Students in the graduate level National Security Studies program at CSUSB, students in the undergraduate anthropology major at UMW, undergraduate and graduate students in Information Sciences and Technology as well as undergraduate students in Criminology, Law and Society at GMU, and undergraduate and graduate students in Political Science and International Studies, Emergency Management, Interdisciplinary Informatics and Cybersecurity at UNO, will be invited to participate. Students are offered the opportunity to participate in the study in exchange for an independent-study or a special topics course credit and the knowledge and experience gained from the course. Following IRB guidelines, participation will not be required and students will be provided with alternative assignments if they choose to no longer participate in the study.

Students at CSUSB will enroll in SSCI 695-01 for 4-credits, a quarter graduate level Independent Studies course in the National Security Studies or National Cyber Security Studies program. Students at UMW will enroll in URES 197 for 2-credits, a semester undergraduate level independent participatory research class. Students at GMU will enroll in AIT 499/CRIM 490 for 3 credits, a summer special topics course. Students at UNO will enroll in PSCI 4920/8926 for 3 credits, another summer special topics course.

5.2 Recruitment Efforts

Recruitment for the study will occur via email advertisements to students enrolled in relevant programs at California State University at San Bernardino (CSUSB), George Mason University (GMU), University of Nebraska at Omaha (UNO), and University of Mary Washington (UMW).

5.3 Incentives for Participation

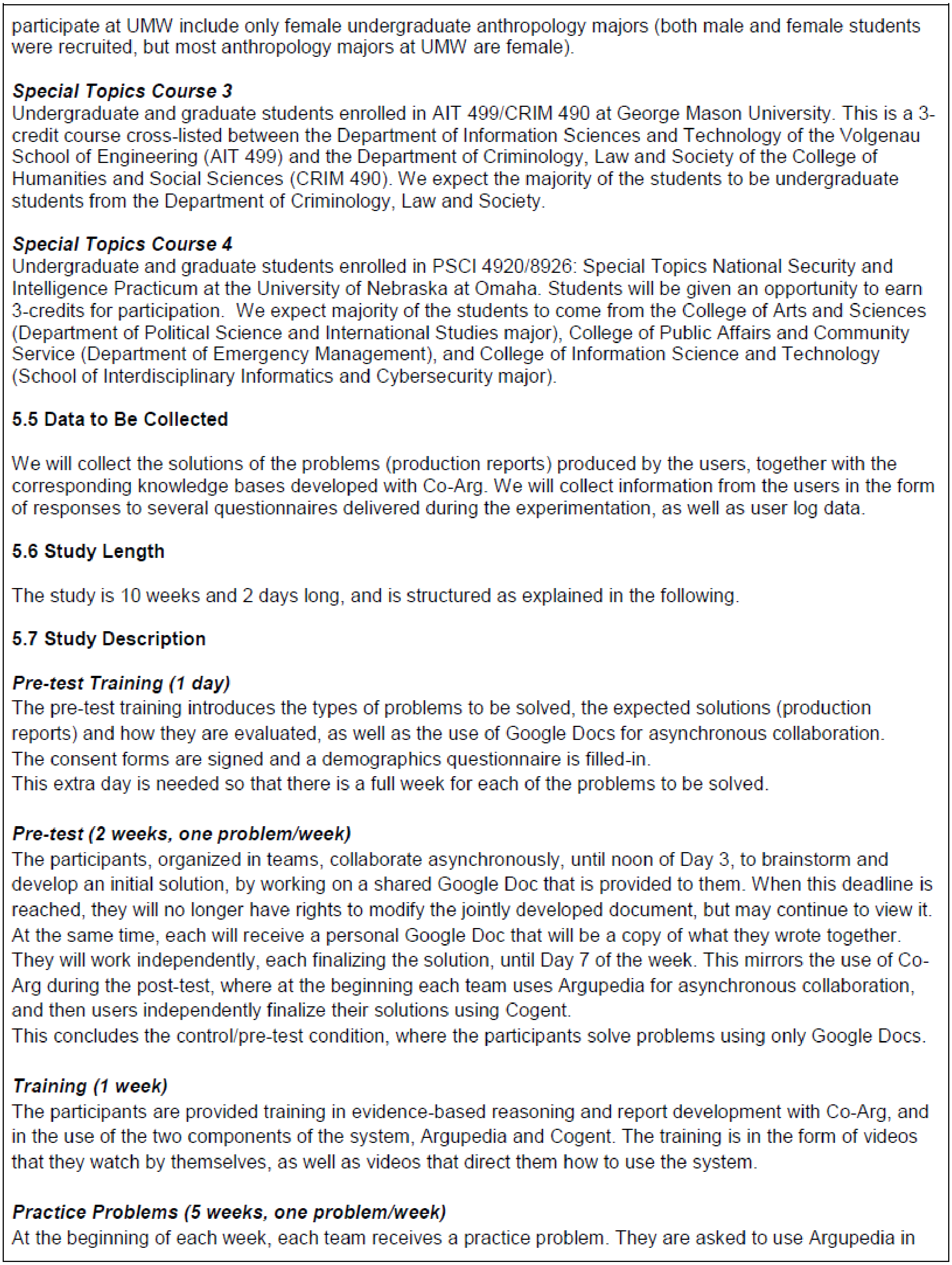
Students are offered the opportunity to participate in the study in exchange for course credit and the knowledge and experience gained from the course. As a result of participation in the study students may improve their quality of reasoning and skill in argumentation. All participants who complete the pre-testing problems, the practice problems, and the post-test problems will be issued a certificate of participation.

5.4 How Subjects are Selected for Eligibility

All subjects selected for participation in this study must be enrolled in either SSCI 695-01 at SCUSB, or in AIT 499/CRIM 490 at GMU, or in PSCI 4920/8926 at UNO, or in URES 197 at UMW.

Independent Study Course 1:
Graduate students enrolled in the National Security Studies Program California State University at San Bernardino (CSUSB). Students will be offered the opportunity to earn 4-credits for participation in the study if they enroll in SSCI 695-01. Currently identified CSUSB students for this study include 20 graduate students, 18 of whom are enrolled in the Master of Arts in National Security Studies, and 2 of whom are enrolled in the Master of Science in National Cyber Security Studies. Of these 20 students, 8 are females and 12 are males.

Independent Study Course 2:
Undergraduate students at University of Mary Washington (UMW). Students will be offered the opportunity to earn 2-credits for participation in the study if they enroll in URES 197. The students who have elected to



participate at UMW include only female undergraduate anthropology majors (both male and female students were recruited, but most anthropology majors at UMW are female).

Special Topics Course 3
Undergraduate and graduate students enrolled in AIT 499/CRIM 490 at George Mason University. This is a 3-credit course cross-listed between the Department of Information Sciences and Technology of the Volgenau School of Engineering (AIT 499) and the Department of Criminology, Law and Society of the College of Humanities and Social Sciences (CRIM 490). We expect the majority of the students to be undergraduate students from the Department of Criminology, Law and Society.

Special Topics Course 4
Undergraduate and graduate students enrolled in PSCI 4920/8926: Special Topics National Security and Intelligence Practicum at the University of Nebraska at Omaha. Students will be given an opportunity to earn 3-credits for participation. We expect majority of the students to come from the College of Arts and Sciences (Department of Political Science and International Studies major), College of Public Affairs and Community Service (Department of Emergency Management), and College of Information Science and Technology (School of Interdisciplinary Informatics and Cybersecurity major).

5.5 Data to Be Collected

We will collect the solutions of the problems (production reports) produced by the users, together with the corresponding knowledge bases developed with Co-Arg. We will collect information from the users in the form of responses to several questionnaires delivered during the experimentation, as well as user log data.

5.6 Study Length

The study is 10 weeks and 2 days long, and is structured as explained in the following.

5.7 Study Description

Pre-test Training (1 day)
The pre-test training introduces the types of problems to be solved, the expected solutions (production reports) and how they are evaluated, as well as the use of Google Docs for asynchronous collaboration.
The consent forms are signed and a demographics questionnaire is filled-in.
This extra day is needed so that there is a full week for each of the problems to be solved.

Pre-test (2 weeks, one problem/week)
The participants, organized in teams, collaborate asynchronously, until noon of Day 3, to brainstorm and develop an initial solution, by working on a shared Google Doc that is provided to them. When this deadline is reached, they will no longer have rights to modify the jointly developed document, but may continue to view it. At the same time, each will receive a personal Google Doc that will be a copy of what they wrote together. They will work independently, each finalizing the solution, until Day 7 of the week. This mirrors the use of Co-Arg during the post-test, where at the beginning each team uses Argupedia for asynchronous collaboration, and then users independently finalize their solutions using Cogent.
This concludes the control/pre-test condition, where the participants solve problems using only Google Docs.

Training (1 week)
The participants are provided training in evidence-based reasoning and report development with Co-Arg, and in the use of the two components of the system, Argupedia and Cogent. The training is in the form of videos that they watch by themselves, as well as videos that direct them how to use the system.

Practice Problems (5 weeks, one problem/week)
At the beginning of each week, each team receives a practice problem. They are asked to use Argupedia in

each of the first two days to contribute to the development of the informal analysis and review the contributions of the other members of the team. At the beginning of Day 3 each user reviews the informal analysis developed by the team and imports the desired results into Cogent. Each user, individually, continues the development of the solution in Cogent, develops the production report, and submits it by the end of Day 6. During Day 7 they receive the solution, an explanation of the solution, and an evaluation grid, and are asked to study them. During a group meeting, they discuss the solution with the instructor.

Post-test (2 weeks, one problem/week)

Each week each team is given a post-test problem to solve with Co-Arg. The process is similar to that from the pre-test, except that they use Co-Arg to solve the problems. During each of the first two days they use Argupedia to contribute to the development of the informal analysis and review the contributions of the other members of the team. At the beginning of Day 3 each user reviews the informal analysis developed as a team and imports the desired results into Cogent. Each user, individually, continues the development of the solution in Cogent, develops the production report, and submits it by the end of Day 7.
This concludes the treatment/post-test condition.

Final Discussion (1 day)

A group meeting during which the users fill-in several questionnaires and discuss the experiment with the instructor.

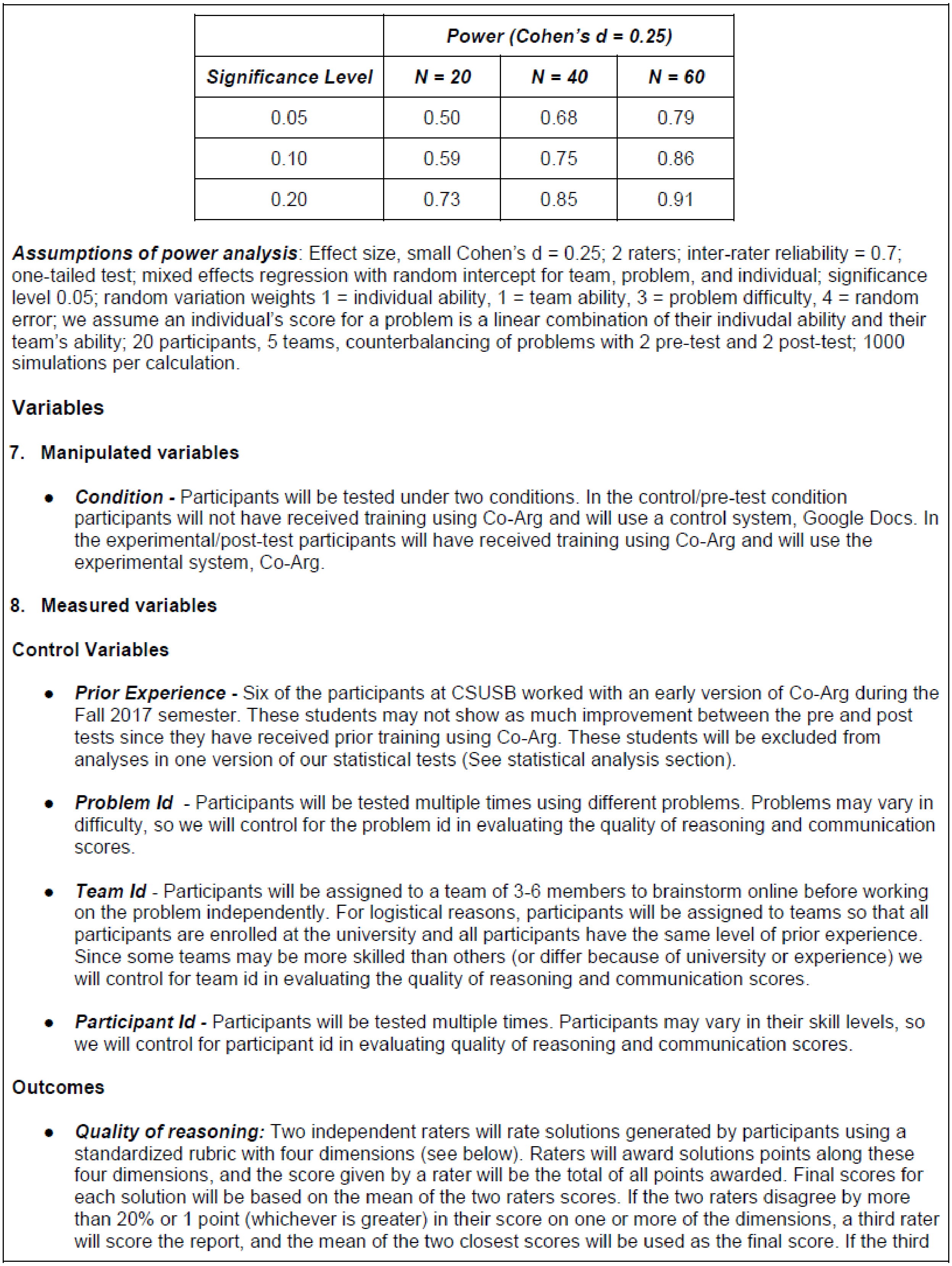
Note: Only the solutions of the pre-test and the post-test problems will be evaluated and compared to determine whether there is improvement in the quality of reasoning and communication. *These solutions will be evaluated together at the end of the experiment.*

6. Sample size

We expect to be able to recruit between 20 and 60 participants (final number will depend on enrollment). While participants will brainstorm online with a small team of 3-6 members at the beginning of each problem, all participants will develop a report individually and their quality of reasoning and communication will be assessed individually. Thus, we consider an individual to be our primary unit of analysis for this study.

We will use ***within-subject design*** in which each participant will develop reports for 4 intelligence problems. 2 problems will be in pre-test and 2 problems will be in post-test. Thus, in total we will have between 80 - 240 observations, 40 - 120 in pre-test condition, 40 - 120 in post-test condition. These observations will come from between 20 - 60 participants (grouped into 5 - 20 brainstorming teams).

The number of individual participants is constrained by the availability and interest of students at the universities we are recruiting from. The number of problems is constrained by the length of the study, which is determined by the length of the term at each university. We anticipate that we will have good power to detect a small effect (Cohen's d = 0.25) if we can enroll 60 participants at a significance level 0.05. If we are only able to enroll around 40 participants we will have modest power at significance level of 0.05 and good power at significance level of 0.20. If are only able to enroll 20 participants we will have only modest power. We conducted power analysis that used the study design, sample size, and statistical analyses as written in this registration (see supplementary R code). We assumed that an individual's score for a problem will be a result of their individual's ability and their team's ability and would also depend on the problem difficulty. Unfortunately, with no prior pilot data we made assumptions about how much each of these factors would contribute to a score. In practice the relative importance of each factor and how the factors interact with each other may be very different from our assumptions and may affect power; thus our power analysis is speculative and not definitive. See results of power analysis below in the table. Results for all 4 problems will be used in confirmatory analyses.



	Power (Cohen's d = 0.25)		
Significance Level	***N = 20***	***N = 40***	***N = 60***
0.05	0.50	0.68	0.79
0.10	0.59	0.75	0.86
0.20	0.73	0.85	0.91

Assumptions of power analysis: Effect size, small Cohen's d = 0.25; 2 raters; inter-rater reliability = 0.7; one-tailed test; mixed effects regression with random intercept for team, problem, and individual; significance level 0.05; random variation weights 1 = individual ability, 1 = team ability, 3 = problem difficulty, 4 = random error; we assume an individual's score for a problem is a linear combination of their indivudal ability and their team's ability; 20 participants, 5 teams, counterbalancing of problems with 2 pre-test and 2 post-test; 1000 simulations per calculation.

Variables

7. Manipulated variables

- ***Condition -*** Participants will be tested under two conditions. In the control/pre-test condition participants will not have received training using Co-Arg and will use a control system, Google Docs. In the experimental/post-test participants will have received training using Co-Arg and will use the experimental system, Co-Arg.

8. Measured variables

Control Variables

- ***Prior Experience -*** Six of the participants at CSUSB worked with an early version of Co-Arg during the Fall 2017 semester. These students may not show as much improvement between the pre and post tests since they have received prior training using Co-Arg. These students will be excluded from analyses in one version of our statistical tests (See statistical analysis section).

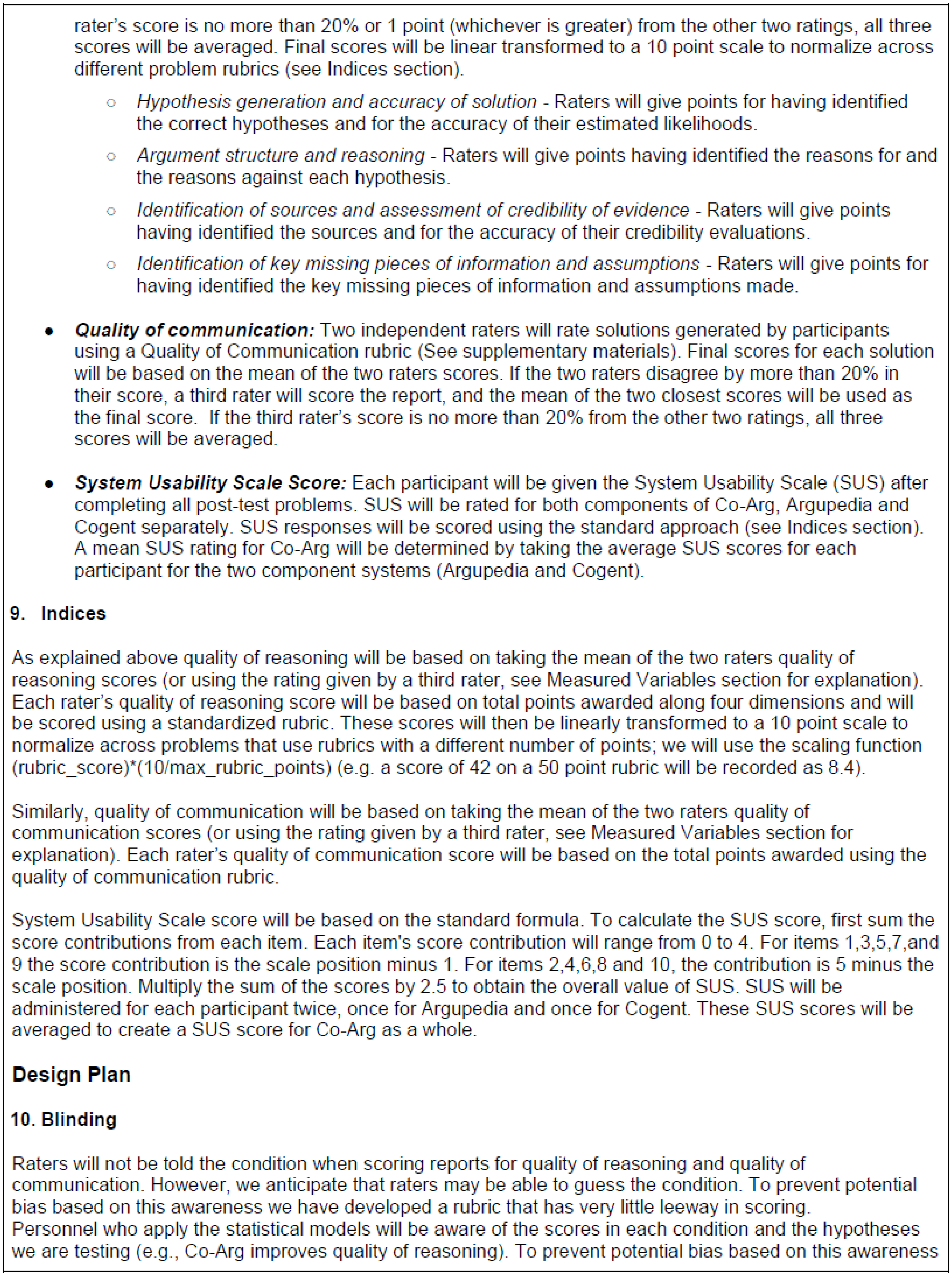
- ***Problem Id*** - Participants will be tested multiple times using different problems. Problems may vary in difficulty, so we will control for the problem id in evaluating the quality of reasoning and communication scores.

- ***Team Id*** - Participants will be assigned to a team of 3-6 members to brainstorm online before working on the problem independently. For logistical reasons, participants will be assigned to teams so that all participants are enrolled at the university and all participants have the same level of prior experience. Since some teams may be more skilled than others (or differ because of university or experience) we will control for team id in evaluating the quality of reasoning and communication scores.

- ***Participant Id -*** Participants will be tested multiple times. Participants may vary in their skill levels, so we will control for participant id in evaluating quality of reasoning and communication scores.

Outcomes

- ***Quality of reasoning:*** Two independent raters will rate solutions generated by participants using a standardized rubric with four dimensions (see below). Raters will award solutions points along these four dimensions, and the score given by a rater will be the total of all points awarded. Final scores for each solution will be based on the mean of the two raters scores. If the two raters disagree by more than 20% or 1 point (whichever is greater) in their score on one or more of the dimensions, a third rater will score the report, and the mean of the two closest scores will be used as the final score. If the third



rater's score is no more than 20% or 1 point (whichever is greater) from the other two ratings, all three scores will be averaged. Final scores will be linear transformed to a 10 point scale to normalize across different problem rubrics (see Indices section).

- ○ *Hypothesis generation and accuracy of solution* - Raters will give points for having identified the correct hypotheses and for the accuracy of their estimated likelihoods.
- ○ *Argument structure and reasoning* - Raters will give points having identified the reasons for and the reasons against each hypothesis.
- ○ *Identification of sources and assessment of credibility of evidence* - Raters will give points having identified the sources and for the accuracy of their credibility evaluations.
- ○ *Identification of key missing pieces of information and assumptions* - Raters will give points for having identified the key missing pieces of information and assumptions made.

- ***Quality of communication:*** Two independent raters will rate solutions generated by participants using a Quality of Communication rubric (See supplementary materials). Final scores for each solution will be based on the mean of the two raters scores. If the two raters disagree by more than 20% in their score, a third rater will score the report, and the mean of the two closest scores will be used as the final score. If the third rater's score is no more than 20% from the other two ratings, all three scores will be averaged.

- ***System Usability Scale Score:*** Each participant will be given the System Usability Scale (SUS) after completing all post-test problems. SUS will be rated for both components of Co-Arg, Argupedia and Cogent separately. SUS responses will be scored using the standard approach (see Indices section). A mean SUS rating for Co-Arg will be determined by taking the average SUS scores for each participant for the two component systems (Argupedia and Cogent).

9. Indices

As explained above quality of reasoning will be based on taking the mean of the two raters quality of reasoning scores (or using the rating given by a third rater, see Measured Variables section for explanation). Each rater's quality of reasoning score will be based on total points awarded along four dimensions and will be scored using a standardized rubric. These scores will then be linearly transformed to a 10 point scale to normalize across problems that use rubrics with a different number of points; we will use the scaling function (rubric_score)*(10/max_rubric_points) (e.g. a score of 42 on a 50 point rubric will be recorded as 8.4).

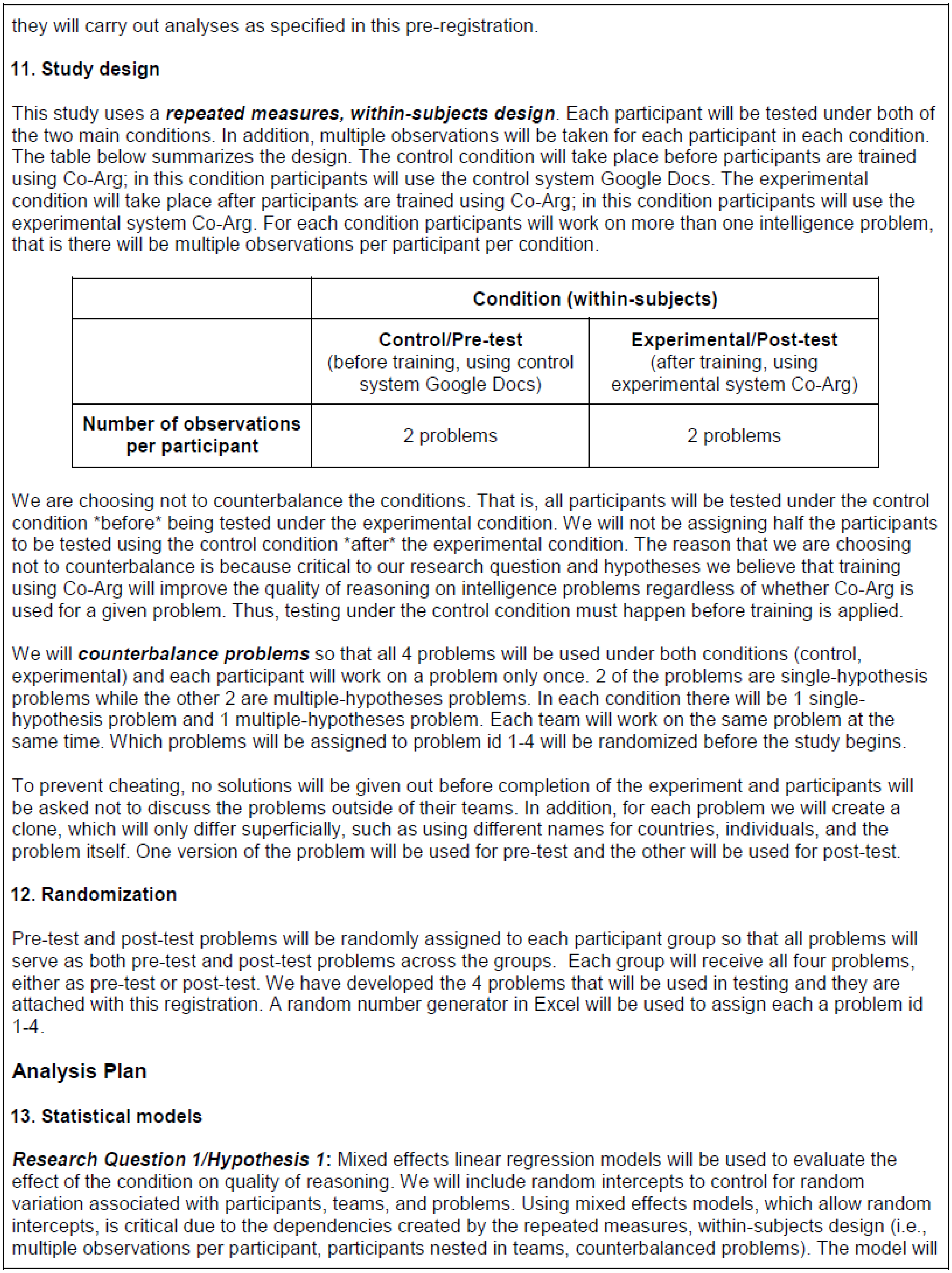
Similarly, quality of communication will be based on taking the mean of the two raters quality of communication scores (or using the rating given by a third rater, see Measured Variables section for explanation). Each rater's quality of communication score will be based on the total points awarded using the quality of communication rubric.

System Usability Scale score will be based on the standard formula. To calculate the SUS score, first sum the score contributions from each item. Each item's score contribution will range from 0 to 4. For items 1,3,5,7,and 9 the score contribution is the scale position minus 1. For items 2,4,6,8 and 10, the contribution is 5 minus the scale position. Multiply the sum of the scores by 2.5 to obtain the overall value of SUS. SUS will be administered for each participant twice, once for Argupedia and once for Cogent. These SUS scores will be averaged to create a SUS score for Co-Arg as a whole.

Design Plan

10. Blinding

Raters will not be told the condition when scoring reports for quality of reasoning and quality of communication. However, we anticipate that raters may be able to guess the condition. To prevent potential bias based on this awareness we have developed a rubric that has very little leeway in scoring.
Personnel who apply the statistical models will be aware of the scores in each condition and the hypotheses we are testing (e.g., Co-Arg improves quality of reasoning). To prevent potential bias based on this awareness



they will carry out analyses as specified in this pre-registration.

11. Study design

This study uses a ***repeated measures, within-subjects design***. Each participant will be tested under both of the two main conditions. In addition, multiple observations will be taken for each participant in each condition. The table below summarizes the design. The control condition will take place before participants are trained using Co-Arg; in this condition participants will use the control system Google Docs. The experimental condition will take place after participants are trained using Co-Arg; in this condition participants will use the experimental system Co-Arg. For each condition participants will work on more than one intelligence problem, that is there will be multiple observations per participant per condition.

	Condition (within-subjects)	
	Control/Pre-test (before training, using control system Google Docs)	**Experimental/Post-test** (after training, using experimental system Co-Arg)
Number of observations per participant	2 problems	2 problems

We are choosing not to counterbalance the conditions. That is, all participants will be tested under the control condition *before* being tested under the experimental condition. We will not be assigning half the participants to be tested using the control condition *after* the experimental condition. The reason that we are choosing not to counterbalance is because critical to our research question and hypotheses we believe that training using Co-Arg will improve the quality of reasoning on intelligence problems regardless of whether Co-Arg is used for a given problem. Thus, testing under the control condition must happen before training is applied.

We will ***counterbalance problems*** so that all 4 problems will be used under both conditions (control, experimental) and each participant will work on a problem only once. 2 of the problems are single-hypothesis problems while the other 2 are multiple-hypotheses problems. In each condition there will be 1 single-hypothesis problem and 1 multiple-hypotheses problem. Each team will work on the same problem at the same time. Which problems will be assigned to problem id 1-4 will be randomized before the study begins.

To prevent cheating, no solutions will be given out before completion of the experiment and participants will be asked not to discuss the problems outside of their teams. In addition, for each problem we will create a clone, which will only differ superficially, such as using different names for countries, individuals, and the problem itself. One version of the problem will be used for pre-test and the other will be used for post-test.

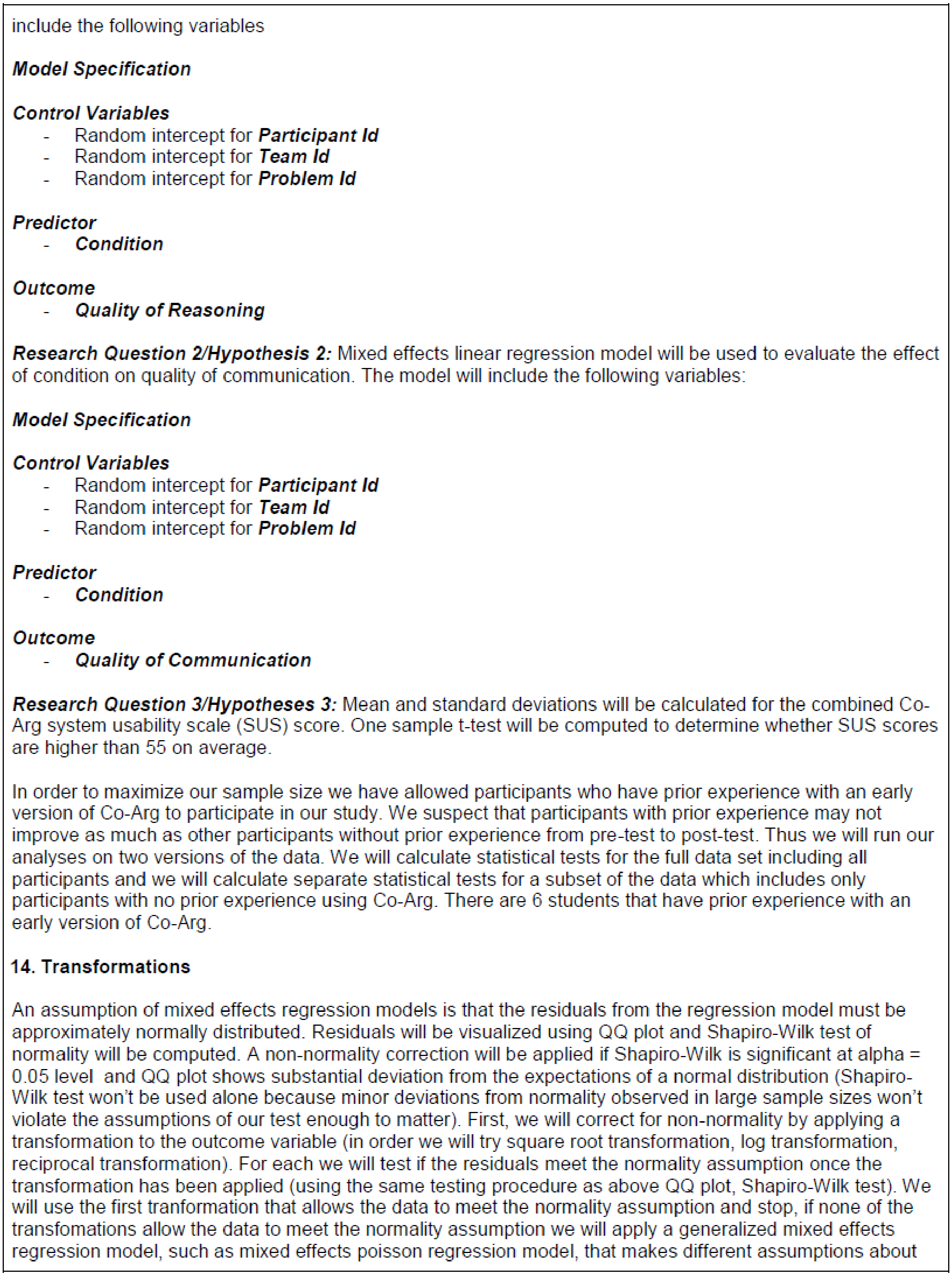
12. Randomization

Pre-test and post-test problems will be randomly assigned to each participant group so that all problems will serve as both pre-test and post-test problems across the groups. Each group will receive all four problems, either as pre-test or post-test. We have developed the 4 problems that will be used in testing and they are attached with this registration. A random number generator in Excel will be used to assign each a problem id 1-4.

Analysis Plan

13. Statistical models

Research Question 1/Hypothesis 1: Mixed effects linear regression models will be used to evaluate the effect of the condition on quality of reasoning. We will include random intercepts to control for random variation associated with participants, teams, and problems. Using mixed effects models, which allow random intercepts, is critical due to the dependencies created by the repeated measures, within-subjects design (i.e., multiple observations per participant, participants nested in teams, counterbalanced problems). The model will



include the following variables

Model Specification

Control Variables

- Random intercept for ***Participant Id***
- Random intercept for ***Team Id***
- Random intercept for ***Problem Id***

Predictor

- ***Condition***

Outcome

- ***Quality of Reasoning***

Research Question 2/Hypothesis 2: Mixed effects linear regression model will be used to evaluate the effect of condition on quality of communication. The model will include the following variables:

Model Specification

Control Variables

- Random intercept for ***Participant Id***
- Random intercept for ***Team Id***
- Random intercept for ***Problem Id***

Predictor

- ***Condition***

Outcome

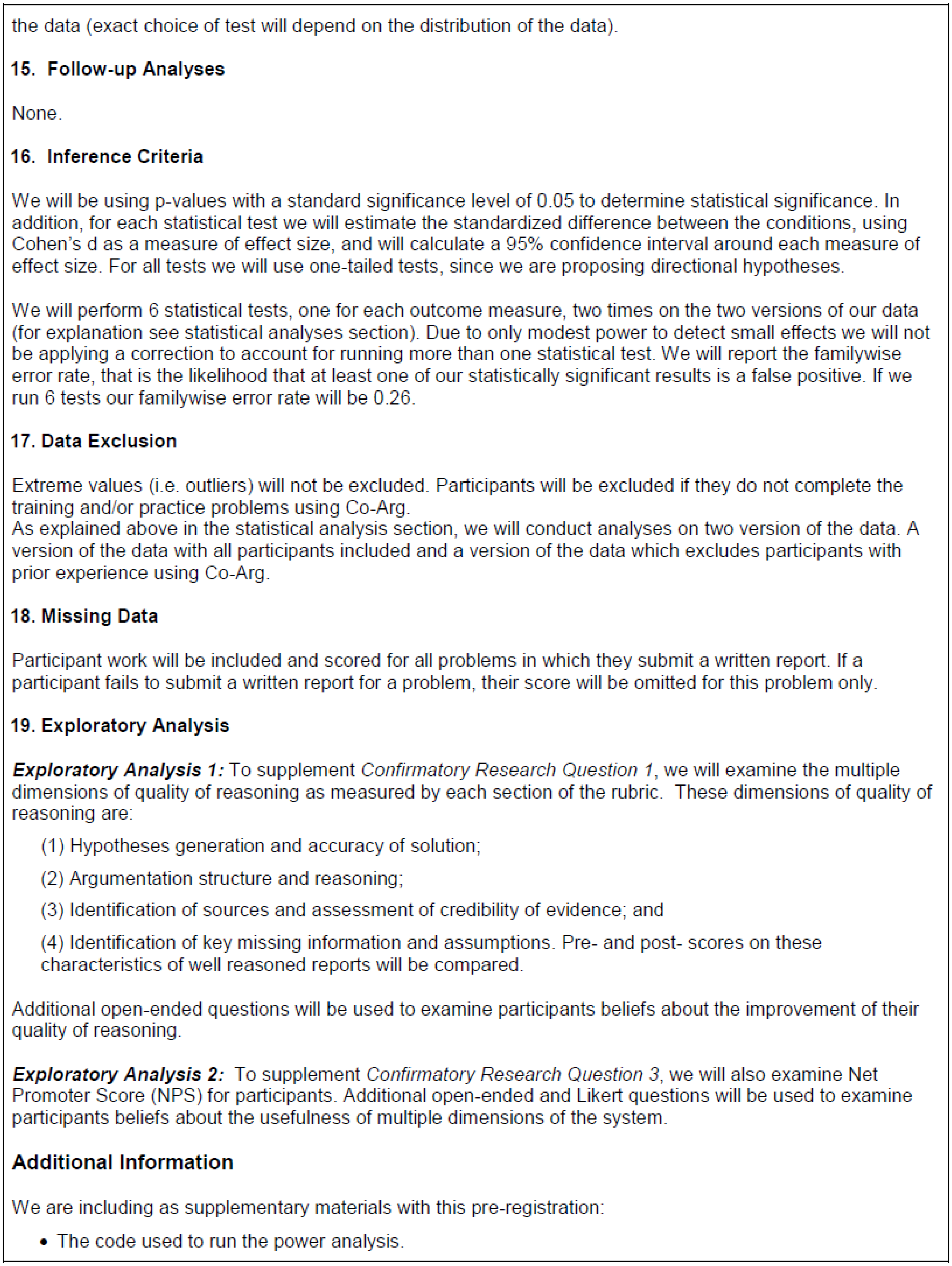
- ***Quality of Communication***

Research Question 3/Hypotheses 3: Mean and standard deviations will be calculated for the combined Co-Arg system usability scale (SUS) score. One sample t-test will be computed to determine whether SUS scores are higher than 55 on average.

In order to maximize our sample size we have allowed participants who have prior experience with an early version of Co-Arg to participate in our study. We suspect that participants with prior experience may not improve as much as other participants without prior experience from pre-test to post-test. Thus we will run our analyses on two versions of the data. We will calculate statistical tests for the full data set including all participants and we will calculate separate statistical tests for a subset of the data which includes only participants with no prior experience using Co-Arg. There are 6 students that have prior experience with an early version of Co-Arg.

14. Transformations

An assumption of mixed effects regression models is that the residuals from the regression model must be approximately normally distributed. Residuals will be visualized using QQ plot and Shapiro-Wilk test of normality will be computed. A non-normality correction will be applied if Shapiro-Wilk is significant at alpha = 0.05 level and QQ plot shows substantial deviation from the expectations of a normal distribution (Shapiro-Wilk test won't be used alone because minor deviations from normality observed in large sample sizes won't violate the assumptions of our test enough to matter). First, we will correct for non-normality by applying a transformation to the outcome variable (in order we will try square root transformation, log transformation, reciprocal transformation). For each we will test if the residuals meet the normality assumption once the transformation has been applied (using the same testing procedure as above QQ plot, Shapiro-Wilk test). We will use the first tranformation that allows the data to meet the normality assumption and stop, if none of the transfomations allow the data to meet the normality assumption we will apply a generalized mixed effects regression model, such as mixed effects poisson regression model, that makes different assumptions about



the data (exact choice of test will depend on the distribution of the data).

15. Follow-up Analyses

None.

16. Inference Criteria

We will be using p-values with a standard significance level of 0.05 to determine statistical significance. In addition, for each statistical test we will estimate the standardized difference between the conditions, using Cohen's d as a measure of effect size, and will calculate a 95% confidence interval around each measure of effect size. For all tests we will use one-tailed tests, since we are proposing directional hypotheses.

We will perform 6 statistical tests, one for each outcome measure, two times on the two versions of our data (for explanation see statistical analyses section). Due to only modest power to detect small effects we will not be applying a correction to account for running more than one statistical test. We will report the familywise error rate, that is the likelihood that at least one of our statistically significant results is a false positive. If we run 6 tests our familywise error rate will be 0.26.

17. Data Exclusion

Extreme values (i.e. outliers) will not be excluded. Participants will be excluded if they do not complete the training and/or practice problems using Co-Arg.
As explained above in the statistical analysis section, we will conduct analyses on two version of the data. A version of the data with all participants included and a version of the data which excludes participants with prior experience using Co-Arg.

18. Missing Data

Participant work will be included and scored for all problems in which they submit a written report. If a participant fails to submit a written report for a problem, their score will be omitted for this problem only.

19. Exploratory Analysis

Exploratory Analysis 1: To supplement *Confirmatory Research Question 1*, we will examine the multiple dimensions of quality of reasoning as measured by each section of the rubric. These dimensions of quality of reasoning are:

(1) Hypotheses generation and accuracy of solution;

(2) Argumentation structure and reasoning;

(3) Identification of sources and assessment of credibility of evidence; and

(4) Identification of key missing information and assumptions. Pre- and post- scores on these characteristics of well reasoned reports will be compared.

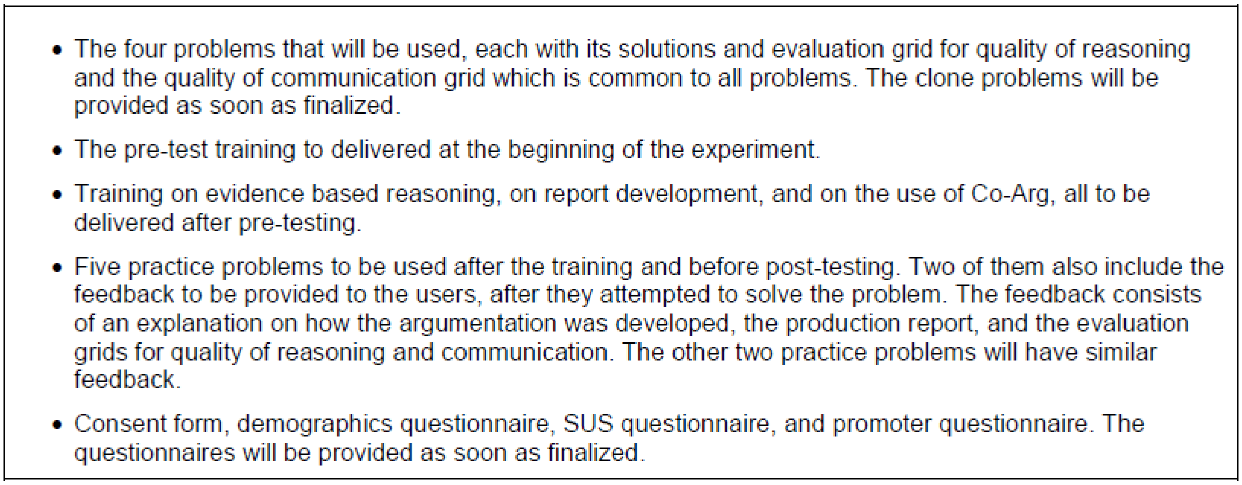
Additional open-ended questions will be used to examine participants beliefs about the improvement of their quality of reasoning.

Exploratory Analysis 2: To supplement *Confirmatory Research Question 3*, we will also examine Net Promoter Score (NPS) for participants. Additional open-ended and Likert questions will be used to examine participants beliefs about the usefulness of multiple dimensions of the system.

Additional Information

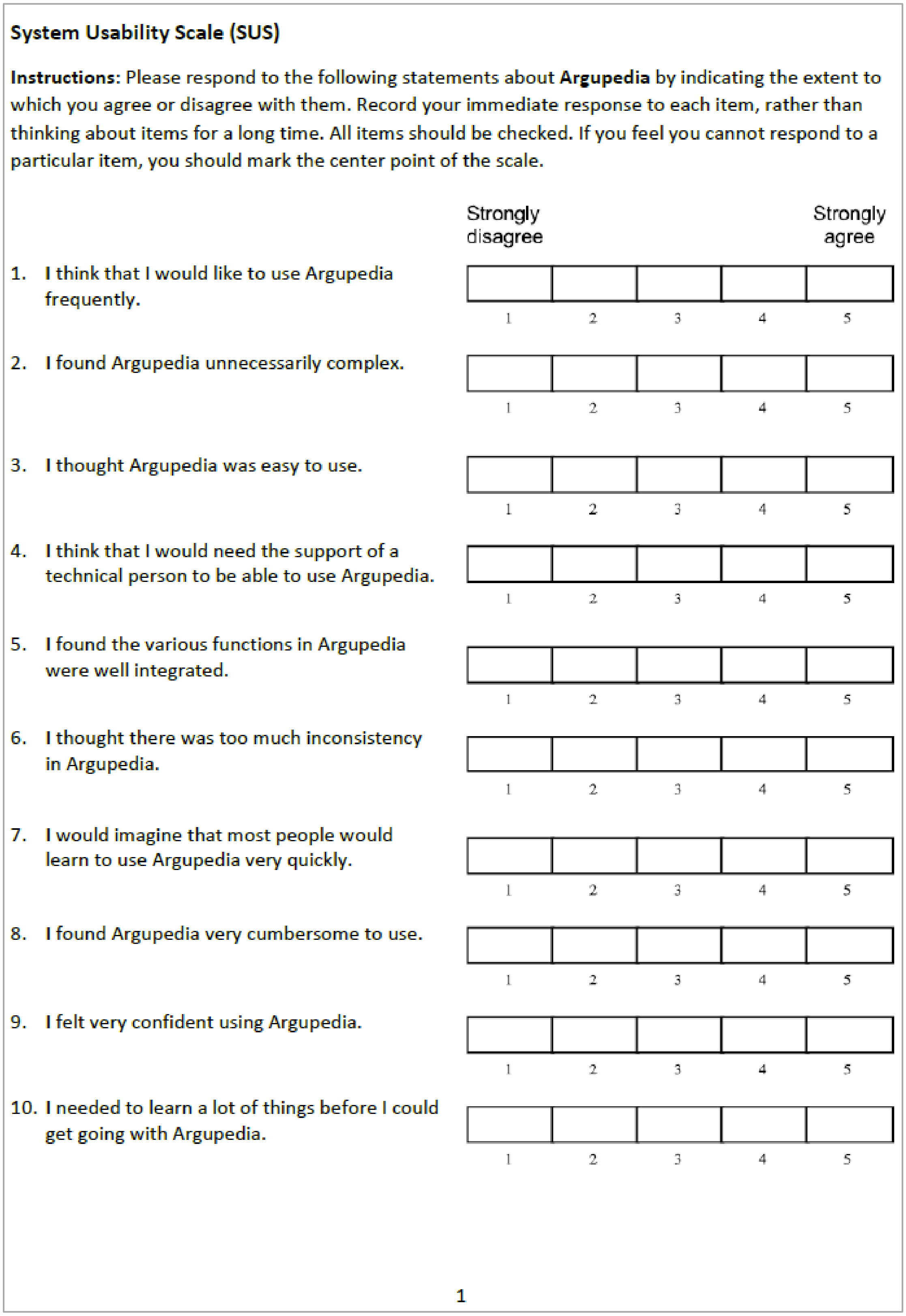
We are including as supplementary materials with this pre-registration:

- The code used to run the power analysis.



- The four problems that will be used, each with its solutions and evaluation grid for quality of reasoning and the quality of communication grid which is common to all problems. The clone problems will be provided as soon as finalized.
- The pre-test training to delivered at the beginning of the experiment.
- Training on evidence based reasoning, on report development, and on the use of Co-Arg, all to be delivered after pre-testing.
- Five practice problems to be used after the training and before post-testing. Two of them also include the feedback to be provided to the users, after they attempted to solve the problem. The feedback consists of an explanation on how the argumentation was developed, the production report, and the evaluation grids for quality of reasoning and communication. The other two practice problems will have similar feedback.
- Consent form, demographics questionnaire, SUS questionnaire, and promoter questionnaire. The questionnaires will be provided as soon as finalized.

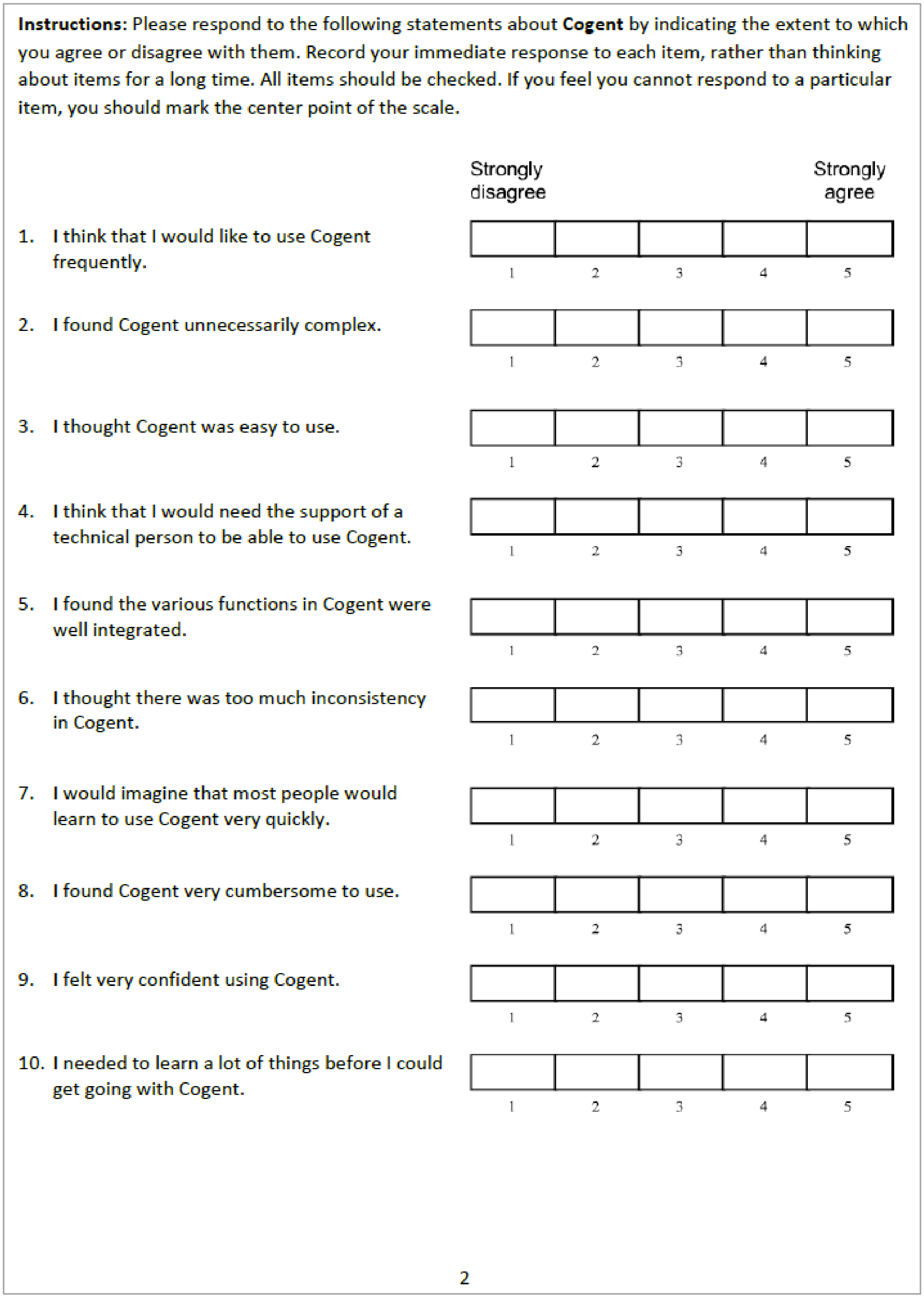
9.2. System Usability Scale



System Usability Scale (SUS)

Instructions: Please respond to the following statements about **Argupedia** by indicating the extent to which you agree or disagree with them. Record your immediate response to each item, rather than thinking about items for a long time. All items should be checked. If you feel you cannot respond to a particular item, you should mark the center point of the scale.

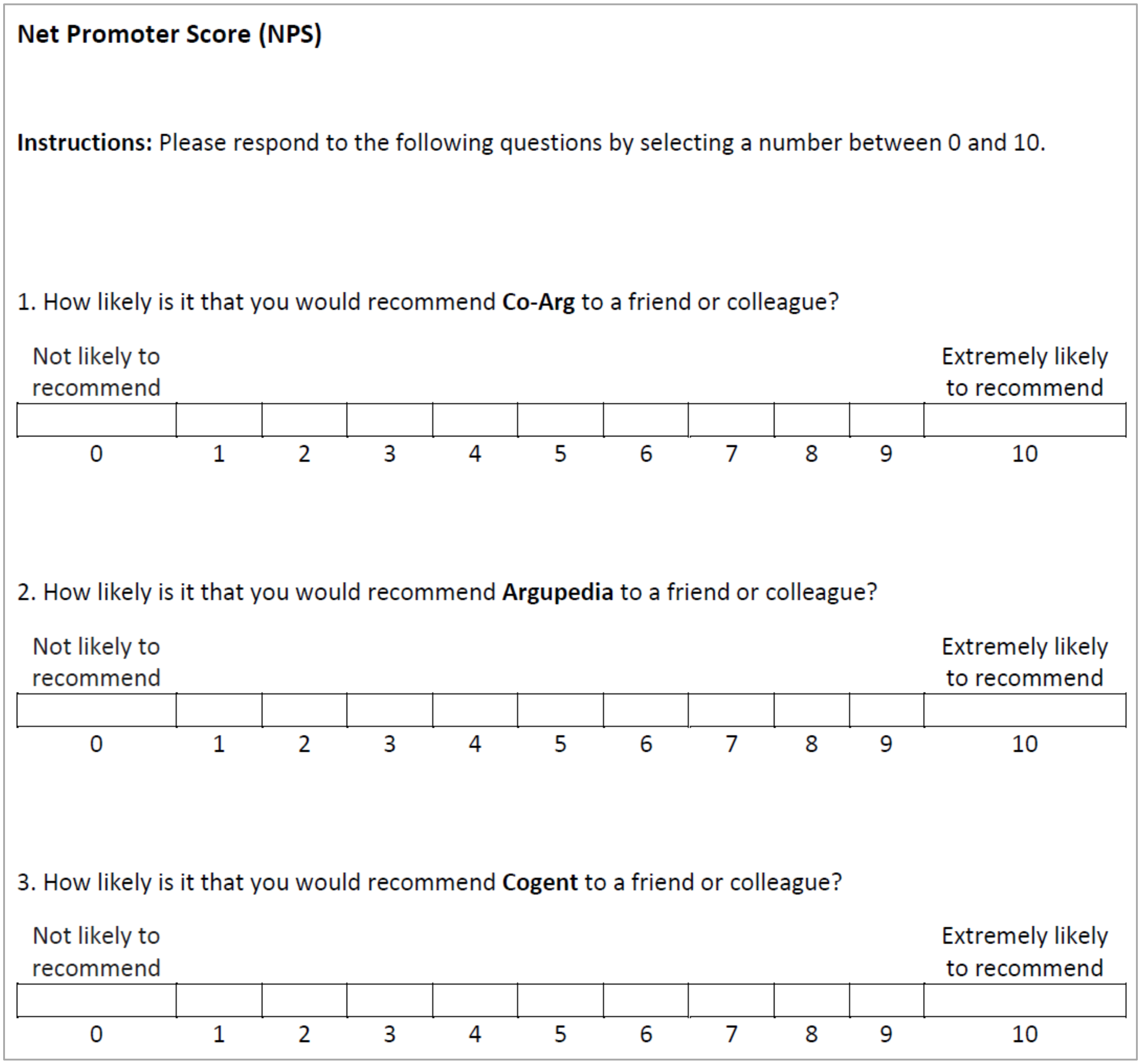
	Strongly disagree				Strongly agree
1. I think that I would like to use Argupedia frequently.	□ 1	□ 2	□ 3	□ 4	□ 5
2. I found Argupedia unnecessarily complex.	□ 1	□ 2	□ 3	□ 4	□ 5
3. I thought Argupedia was easy to use.	□ 1	□ 2	□ 3	□ 4	□ 5
4. I think that I would need the support of a technical person to be able to use Argupedia.	□ 1	□ 2	□ 3	□ 4	□ 5
5. I found the various functions in Argupedia were well integrated.	□ 1	□ 2	□ 3	□ 4	□ 5
6. I thought there was too much inconsistency in Argupedia.	□ 1	□ 2	□ 3	□ 4	□ 5
7. I would imagine that most people would learn to use Argupedia very quickly.	□ 1	□ 2	□ 3	□ 4	□ 5
8. I found Argupedia very cumbersome to use.	□ 1	□ 2	□ 3	□ 4	□ 5
9. I felt very confident using Argupedia.	□ 1	□ 2	□ 3	□ 4	□ 5
10. I needed to learn a lot of things before I could get going with Argupedia.	□ 1	□ 2	□ 3	□ 4	□ 5



Instructions: Please respond to the following statements about **Cogent** by indicating the extent to which you agree or disagree with them. Record your immediate response to each item, rather than thinking about items for a long time. All items should be checked. If you feel you cannot respond to a particular item, you should mark the center point of the scale.

	Strongly disagree				Strongly agree
1. I think that I would like to use Cogent frequently.	□ 1	□ 2	□ 3	□ 4	□ 5
2. I found Cogent unnecessarily complex.	□ 1	□ 2	□ 3	□ 4	□ 5
3. I thought Cogent was easy to use.	□ 1	□ 2	□ 3	□ 4	□ 5
4. I think that I would need the support of a technical person to be able to use Cogent.	□ 1	□ 2	□ 3	□ 4	□ 5
5. I found the various functions in Cogent were well integrated.	□ 1	□ 2	□ 3	□ 4	□ 5
6. I thought there was too much inconsistency in Cogent.	□ 1	□ 2	□ 3	□ 4	□ 5
7. I would imagine that most people would learn to use Cogent very quickly.	□ 1	□ 2	□ 3	□ 4	□ 5
8. I found Cogent very cumbersome to use.	□ 1	□ 2	□ 3	□ 4	□ 5
9. I felt very confident using Cogent.	□ 1	□ 2	□ 3	□ 4	□ 5
10. I needed to learn a lot of things before I could get going with Cogent.	□ 1	□ 2	□ 3	□ 4	□ 5

9.3. Net Promoter Score



Net Promoter Score (NPS)

Instructions: Please respond to the following questions by selecting a number between 0 and 10.

1. How likely is it that you would recommend **Co-Arg** to a friend or colleague?

Not likely to recommend

Extremely likely to recommend

0	1	2	3	4	5	6	7	8	9	10

2. How likely is it that you would recommend **Argupedia** to a friend or colleague?

Not likely to recommend

Extremely likely to recommend

0	1	2	3	4	5	6	7	8	9	10

3. How likely is it that you would recommend **Cogent** to a friend or colleague?

Not likely to recommend

Extremely likely to recommend

0	1	2	3	4	5	6	7	8	9	10

9.4. Quality of Communication Evaluation Grid

Evaluation Rubric	
Quality of communication	**Points**
Main conclusion is clearly stated up front, **and** strong organization and writing allow the reader to quickly identify the main reasons in the argument that favor or disfavor the main conclusion. The analysis that connects the evidence to these reasons is presented clearly and precisely. As a result, the reader quickly understands the argumentation supporting the reasons and the significance of individual items of evidence. The evidence is portrayed accurately and clearly.	6
Main conclusion is stated up front, **and** a mostly coherent organization includes numerous examples of cohesive and clear ideas. The reasons favoring or disfavoring the main conclusion are apparent to the reader. The underlying analysis explaining the basis for these reasons is mostly comprehensible but garbled and confusing in a few places. As a result, the reader does not quickly understand all the argumentation supporting the reasons. A few details appear disjointed and the relevance of some evidence is not clear. The description of a few items of evidence is not clear.	4
Main conclusion is stated but is not up front, **or** the presentation of ideas lacks coherence and cohesiveness. The organization is somewhat muddled (due in part to repetition) and the reader has difficulty quickly discerning the reasons favoring or disfavoring the main conclusion (due in part to statements that are contradictory). The underlying analysis explaining the basis for these reasons is murky and not easily comprehended in several places. As a result, the reader does not quickly understand much of the argumentation supporting the reasons. Numerous details appear disjointed and the relevance of many items of evidence is not clear. The description of many items of evidence is not clear.	2
The report is largely incomprehensible **and** lacks a coherent organization. The main reasons favoring or disfavoring the main conclusion are not apparent to the reader, either because these reasons are not adequately linked back to the main conclusion or the writing cannot be easily comprehended. The analysis discussing the evidence is extensively garbled and confusing. The language describing the evidence is similarly confusing.	0

9.5. Sample Quality of Reasoning Evaluation Grid with General Scoring Guidance

Accuracy of Solution	• Answers the intelligence question • Provides the expected solution • Probabilistic assessment of the key judgment is consistent with the underlying uncertainties
Argument Structure and Reasoning	• Reasoning connecting sub-judgments is clear and supports (favors) the key judgment • Reasoning connecting evidence and assumptions to judgments is clear and supports the judgments • Reasoning is free of logical errors and biases • Reasoning accurately assesses relevance of key evidence for expected conclusion
Evidence Assessment	• Uses all relevant information, both favoring and disfavoring • Identifies the sources of information and accurately describes the attributes reflecting the assessed credibility of this evidence, especially for key evidence
Uncertainty	• Identifies key evidence and missing information that that result in uncertainty, and the impact that uncertainty has on their judgments • Probabilistic assessments of the main judgment and underlying sub-judgments are consistent • Reasoning and evidence assessments support the probability expressed in the judgments
Assumptions (if applicable)	• Identifies information gaps and assumptions required to support relevant judgments • Provides and justifies the assessed probabilities of the assumptions • Identifies key assumptions and new information that could change the main judgment
Analysis of Alternatives (if applicable)	• Identifies and ranks expected alternative hypotheses • Considers both favoring and disfavoring evidence and reasons for the preferred hypothesis and alternatives

Table 34: Quality of Reasoning Dimensions in the Evaluation Grid

Criteria	**Scoring Guidance**
Argument Structure and Reasoning • Reasoning connecting sub-judgments is clear and supports (favors) the key judgment • Reasoning connecting evidence and assumptions to judgments is clear and supports the judgments • Reasoning is free of logical errors and biases	0 (Poor)—Lacks a coherent structure that links the evidence to the sub-judgments and the key judgment 1 (Fair)—Consists largely of assertions with few explanations of the reasoning connecting evidence to sub-judgments and sub-judgments to the key judgment; contains clear logical errors or biases undermining the reasoning 2 (Satisfactory)—Offers some explanations but reasoning linking evidence to sub-judgments and the key judgment not consistently clear; reflects some logical error or biases that weaken the reasoning 3 (Good)—Has a coherent structure that generally links the sub-judgments to the key judgment and the evidence to the sub-judgments with few if any logical errors 4 (Excellent)—Has a coherent structure that offers sophisticated reasoning linking evidence to sub-judgments and sub-judgments to the key judgment and is free of logical errors or biases
Score:___	
Guidance The conclusion that Manada is selling the Devastator SAM system is based on the reasoning that the SAM tested on 24 June—when a delegation from Sindia was to observe a test of the SAM it was buying—demonstrated range and target-engagement capabilities that are consistent with the Devastator. The conclusion that Manada is not selling the Demolisher or Destructor SAM is based on the reasoning that the SAM tested on 24 June demonstrated target-engagement capabilities that are not consistent with either the Demolisher or the currently available version of the Destructor.	

Table 35: Sample Qualitative Assessment with Specific Guidance

9.6. Sample Quality of Reasoning Evaluation Grid with Problem-Specific Scoring Guidance

Which SAM system is Manada selling Sindia?	
Criterion	**Points**
	Total 28
Hypothesis generation and accuracy of solution	**Subtotal: 9**
Addresses each of the following hypotheses: • Manada is selling the Destructor SAM to Sindia. • Manada is selling the Devastator SAM to Sindia. • Manada is selling the Demolisher SAM to Sindia.	3 (1 each)
Solution: • Manada is likely (55-70%) selling the Devastator.	1 point for likely (55-70%); 0.5 for barely likely (50-55%) or more than likely (70-80%); additional 1 point if numerical probability included
• Manada is more than unlikely (20-30%) selling the Destructor.	1 point for more than unlikely (20-30%); 0.5 for very unlikely (5-20%) or unlikely (30-45%); additional 1 point if numerical probability included
• Manada is very unlikely (5-20%) selling the Demolisher.	1 point for very unlikely (5-20%); 0.5 for almost no chance (1-5%) or more than unlikely (20-30%); additional 1 point if numerical probability included

Argument structure and reasoning	**Subtotal: 10**
Reasoning behind Devastator	
• (For) Devastator has the same range as SAM tested on 24 June Reasoning behind Demolisher SAM • (For) Devastator has the same target-engagement altitude of SAM tested on 24 June. Source: intercepted communication and data from missile-tracking radar station	2 (1 each)
Reasoning behind Destructor SAM	
• (For) Destructor has the same target-altitude engagement capability as SAM tested on 24 June. • (For) Manada publicly announced on 1 May 2017 that it was selling Sindia the Destructor SAM • (Against) There were no Destructor SAMs available for testing in June that had a range of 680 km • (Against) A longer range variant of the Destructor SAM with a range of 500 km was not available for testing in June • (Against) Development of the longer-range variant was not to be completed until November 2017. • (Against) Funding for the institute developing the longer range variant was not increased	6 (1 each)
Reasoning behind Demolisher SAM	
• (For) Demolisher has the same range as SAM tested on 24 June • (Against) The target-engagement altitude is inconsistent with the SAM tested on 24 June	2 (1 each)

Identification of sources and assessment of credibility of evidence	**Subtotal: 6**
Considers information from technical collection, historical data on SAMs, human sources, and intercepted communications is credible	2
Sourcing:	
• Provides extensive source references	4
• Provides few source references	2
• Provides hardly any source references	0
Identification of key missing information and assumptions	**Subtotal: 3**
No information available on whether longer-range SAM was ready for testing in June; assumption: development of longer range variant could not be accelerated so significantly without increased funding	3

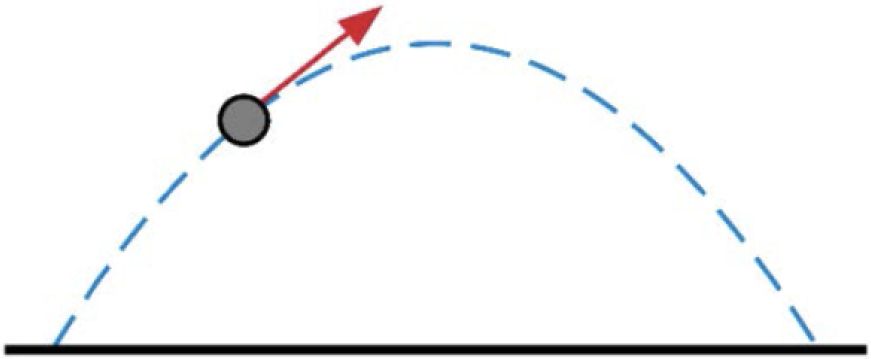
9.7. “Fillistan Conducts Ballistic Missile Tests” Problem

This is a fictional case, not based on real-world events or entities. All the information you need to answer this question is provided along with the problem. The information provided may not be entirely conclusive or comprehensive, and answers may not be absolutely certain. Please do not search for additional information online-- this is a test of reasoning rather than your ability to search for new clues. You should use your best judgment and reasoning in arriving at your answer, explain what alternatives you considered, and give your arguments for why a reader should believe your conclusions.

Fillistan Conducts Ballistic Missile Tests

Scenario

Situation: The countries Fillistan and Buland are strategic enemies. Fillistan has been developing liquid-fueled ballistic missiles for several years now. "Ballistic" missiles are those whose engines operate for a relatively short period of time (compared to the overall flight time) and then shut down. The missile then follows a free-fall parabolic trajectory, just like when you throw a rock a long distance. On 10 January 2017, a missile launch was detected by Buland technical intelligence, originating from an area known to be Fillistan's missile development facility and test range. Ballistic missiles have a maximum range that is determined by the missile's weight, fuel capacity,and the efficiency of its rocket engines. The rocket engines can be shut down early to strike targets that are located at distances less than the maximum range of the missile. Fillistan has or is developing multiple kinds of ballistic missiles, described in more detail below.



Because Buland's security is threatened by Fillistan's missile development, Buland has emphasized intelligence collection on Fillistan's missile development. Their efforts have proved successful, and Buland now has five highly placed human sources reporting on Fillistan's missile development. The sources are managed and tasked out of a single office in Buland's intelligence service. Intelligence officers in this facility have access to identifying information on all five sources.

Buland has also been able to decipher the encryption system for Fillistan's military communications. Fillistan, however, is confident that its communications using this encryption system remain secure.

Fillistan is an authoritarian, highly paranoid state. For security purposes, Fillistan's missile program is highly compartmentalized, which means that individuals working on one missile program will not have knowledge of others and work in different locations. Fillistan provides almost no public information on its missiles until development has been completed and they are deployed.

Key Question

What kind of ballistic missile was launched from Fillistan on 10 January 2017?

Available Information

- A large body of intelligence on Fillistan's ballistic missile capacity has been gathered from multiple sources over time. This information is summarized in Table 1, which shows the main parameters for Fillistan's two known operational liquid-fuel missiles. For the purposes of this problem, you may assume that the information in Table 1 is accurate and has been independently corroborated.

Table 1: Missile Characteristics

Missile Name	Status	Demonstrated Range (km)	Max Enginge Operation Time (s)
Victory	Operational	300-320	60
Revolution	Operational	800-1,000	128

- In contrast with liquid-fuel missiles, the main advantage of solid-fuel missiles is that fuel can safely be stored inside the missiles. Consequently, solid-fuel missiles do not require fueling before launch. Liquid fuel, on the other hand, is particularly unstable and combustible, and must be stored separately from the rest of the missile. A liquid-fuel missile is filled shortly before launch.

- According to an intercept of an encrypted communication, a Fillistan Defense Ministry official in January 2016 told another Defense Ministry official that development of solid-fuel missiles had become a high priority. Furthermore, he stated, funding for the development of solid-fuel missiles had increased by a factor of four, to a total of about $6 billion focused on swift creation of a functional prototype.

- In June 2016, a reliable source in the Best Flight R&D Institute in the city of Drogan reported that Fillistan was designing a new solid-fuel missile, named Progress. At the time of the report, the Institute was about one year away from building and testing a prototype. The source began working for Best Flight after a transfer from another institute.

- In October 2016, a reliable source in the guidance department at the High Flight Guidance R&D Institute at Poolton reported that modifications were being made to the inertial guidance system (a guidance system that uses a computer and motion sensors) of the Revolution missile.

- In October 2016, a new source in Fillistan's General Staff who lacks an established reporting record reported that Fillistan's military and missile designers are completely satisfied with the reliability and capabilities of the Victory missile.

- In late November 2016, the head of the institute responsible for designing inertial guidance systems told Fillistan's Defense Minister that the design changes for the Revolution missile inertial guidance system would be completed by the end of the year. This exchange was captured in an intercept of an encrypted communication.

- According to an intercept of an encrypted message, on 22 December 2016, a Fillistan Defense Ministry official asked Andrew Hall (identified as the lead designer of the Progress missile) when the "solid-fueled missile" would be ready for testing. Hall responded, "in about three months."

- Buland's intelligence community estimates with 70 percent confidence that about $3 billion is needed for Fillistan to develop a new missile.

- In late December 2016, a source who works in Fillistan's Missile Materials R&D Institute was summarily executed, according to an encrypted Interior Ministry communication. The source, according to the intercepted message, refused to provide any information during his interrogation.

- In late December 2016, the source who works for the Best Flight R&D Institute was arrested, according to an intercept of an encrypted Fillistan Interior Ministry communication.

- Fillistan conducted no missile tests during the October-December 2016 period.

- In early January 2017, the source in the guidance department at Fillistan's High Flight Guidance Institute detected Interior Ministry surveillance over a period of several days and requested immediate extraction from Fillistan.

- On 8 January 2017, an engineer at the Best Flight R&D Institute told a colleague that development of the Progress missile was going faster than expected, according to an intercept of an encrypted message.

- On 10 January 2017, the Buland space-based missile warning system detected a missile launch from Fillistan at 0945 local time. Analysis of the missile's signature placed the launch location within a 5.2 km radius of the known launch complex at the Matana missile development and test facility near the capital Umbabwea. (This launch location is unusual for an operational system; it is Fillistan practice to launch operational missiles from exercise areas in the field.) According to heat-signature data from the missile-warning system, the missile's engine operated for 90 seconds. Two days later, the heat-measurement component of the warning system was calibrated. Engineers assessed that prior to calibration, the measurement of burn time was off by as much as 25 percent due to a major programming error. Engineers told missile analysts that they are 75 percent confident that the maximum extent to which data could be off was 25 percent.

- On 10 January 2017, the Observation (a Buland missile-tracking radar ship deployed 150 km off Fillistan's coast detected a missile rising from Fillistan territory. The launch location was derived by back-tracking the trajectory and was assessed to be within a 1.5 km radius of the Matana launch complex. The missile's heading was due west, and flew directly over the ship. The missile impacted the ocean 500 km down-range from the launch point. All radar and tracking systems on the Observation were operational and recently calibrated.

- In late February 2017, the new source in Fillistan's General Staff reported that the missile tested on 10 January was a Victory missile.

- According to an intercept of an encrypted message, on 1 February 2017, Jay Wilson (a senior Fillistan defense official in Umbabwea) called an individual identified as Alan Sampson at an unidentified R&D institute and asked whether he had completed analysis of the missile telemetry data from the 10 January test.

- According to an intercept of encrypted communications, on 1 March 2017, Alan Sampson called an individual identified as Jay Wilson and told him analysis of the telemetry data had been completed and was being reviewed by Mark Pullman, his institute's deputy director. Sampson noted that his institute was happy with the successful launch since all of the institute's efforts had gone into this crash program, the first missile this institute has been tasked with developing. The missile associated with the "crash program" was not identified.

- Andrew Hall was given a national "Hero" award in a highly publicized ceremony in early March 2017 for his accomplishments in "advancing Fillistan's national security."

- On 20 March 2017, according to the Observation and Buland's space-based missile warning system, a flight test of a missile exploded at launch (before clearing the launch tower) at the Matana complex. Overhead photography of the site shows extensive debris scattered around the launch point in a pattern that is consistent with a missile exploding before it cleared the launch tower.

- According to an intercept of encrypted communications, Marie Harris called John Doesky on 21 March 2017 and told him there is "nothing left of the missile," but the debris is being collected and will be sent to his institute at Pamplan. The institute was not identified.

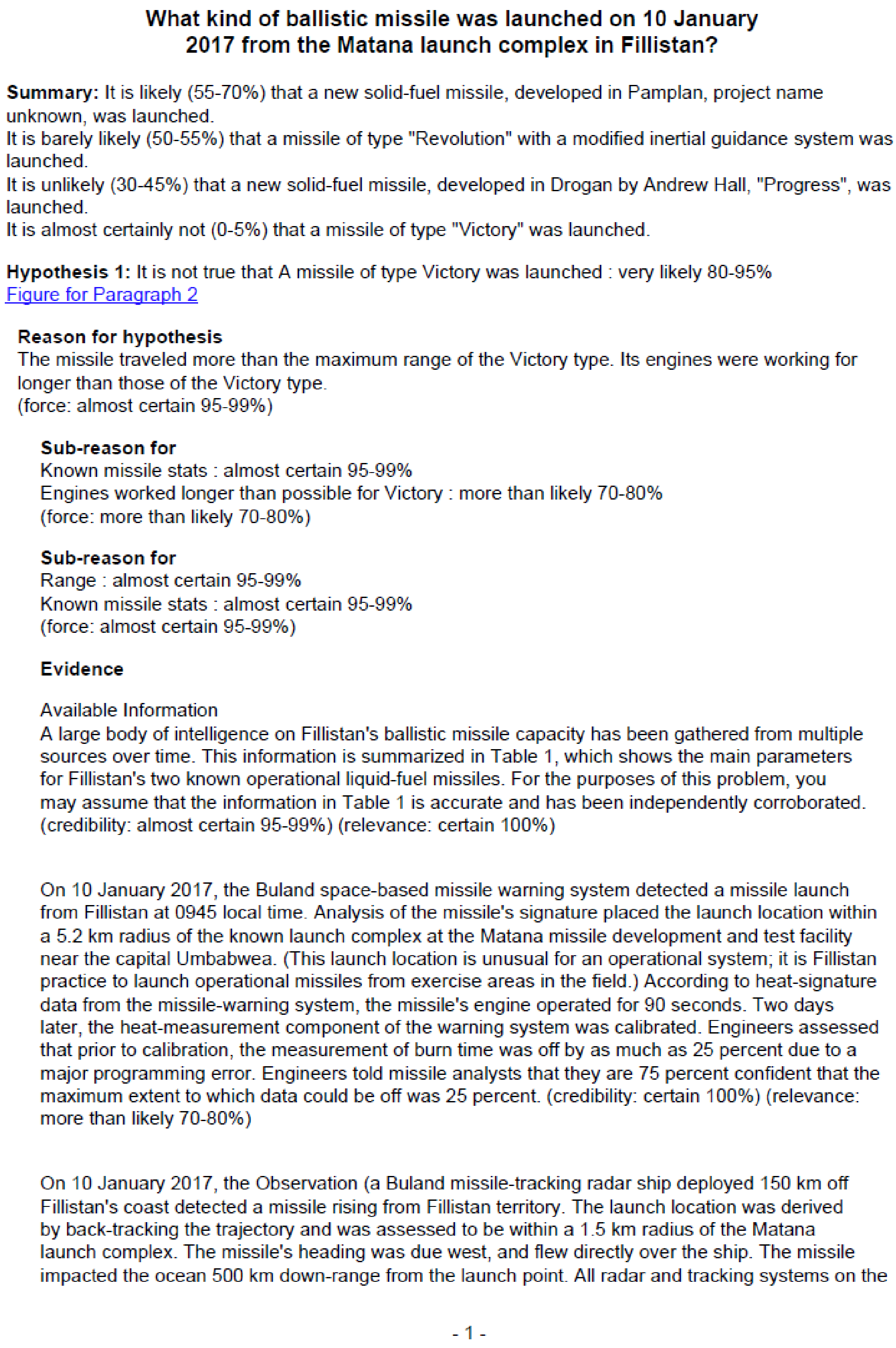
- A reliable source who works in Fillistan's Defense Ministry (the department responsible for all of Fillistan's missile development) reported in late March 2017 that Fillistan did not have any solid-fueled missiles in design or under development. The source also reported that Andrew Hall received the Hero award for his work on the Revolution missile. The source has reported reliably for ten years on a variety of defense issues. The source now works at a remote location and there had been no contact with the source for more than a year. In March 2017, the source travelled to the capital where contact was made.

- According to an intercept of a phone conversation between two missile engineers at Fillistan's GoForSpace Institute, the head of the GoForSpace institute (who designed the Victory missile) was abruptly dismissed on 15 April 2017. Both engineers were perplexed by the dismissal.

- According to an intercept, the secretary of an unidentified R&D institute made airline reservations for an individual identified as Deputy Director Mark Pullman and another individual identified as Director John Doesky to fly from Pamplan to the capital of Fillistan on 30 April.

- Both the Victory and Revolution missiles have been test fired in numerous exercises over the past decade, all of which were successful. The successful launch rate for the Victory missile (including both developmental and exercise launches) is 85 percent. The successful launch rate for the Revolution missile (including both developmental and exercise launches) is 90 percent.

9.8. The Solution “dolphinsfin” of the “Fillistan” Problem in the T&E Experiment



What kind of ballistic missile was launched on 10 January 2017 from the Matana launch complex in Fillistan?

Summary: It is likely (55-70%) that a new solid-fuel missile, developed in Pamplan, project name unknown, was launched.
It is barely likely (50-55%) that a missile of type "Revolution" with a modified inertial guidance system was launched.
It is unlikely (30-45%) that a new solid-fuel missile, developed in Drogan by Andrew Hall, "Progress", was launched.
It is almost certainly not (0-5%) that a missile of type "Victory" was launched.

Hypothesis 1: It is not true that A missile of type Victory was launched : very likely 80-95%
Figure for Paragraph 2

Reason for hypothesis
The missile traveled more than the maximum range of the Victory type. Its engines were working for longer than those of the Victory type.
(force: almost certain 95-99%)

Sub-reason for
Known missile stats : almost certain 95-99%
Engines worked longer than possible for Victory : more than likely 70-80%
(force: more than likely 70-80%)

Sub-reason for
Range : almost certain 95-99%
Known missile stats : almost certain 95-99%
(force: almost certain 95-99%)

Evidence

Available Information
A large body of intelligence on Fillistan's ballistic missile capacity has been gathered from multiple sources over time. This information is summarized in Table 1, which shows the main parameters for Fillistan's two known operational liquid-fuel missiles. For the purposes of this problem, you may assume that the information in Table 1 is accurate and has been independently corroborated. (credibility: almost certain 95-99%) (relevance: certain 100%)

On 10 January 2017, the Buland space-based missile warning system detected a missile launch from Fillistan at 0945 local time. Analysis of the missile's signature placed the launch location within a 5.2 km radius of the known launch complex at the Matana missile development and test facility near the capital Umbabwea. (This launch location is unusual for an operational system; it is Fillistan practice to launch operational missiles from exercise areas in the field.) According to heat-signature data from the missile-warning system, the missile's engine operated for 90 seconds. Two days later, the heat-measurement component of the warning system was calibrated. Engineers assessed that prior to calibration, the measurement of burn time was off by as much as 25 percent due to a major programming error. Engineers told missile analysts that they are 75 percent confident that the maximum extent to which data could be off was 25 percent. (credibility: certain 100%) (relevance: more than likely 70-80%)

On 10 January 2017, the Observation (a Buland missile-tracking radar ship deployed 150 km off Fillistan's coast detected a missile rising from Fillistan territory. The launch location was derived by back-tracking the trajectory and was assessed to be within a 1.5 km radius of the Matana launch complex. The missile's heading was due west, and flew directly over the ship. The missile impacted the ocean 500 km down-range from the launch point. All radar and tracking systems on the

Observation were operational and recently calibrated. (credibility: almost certain 95-99%) (relevance: certain 100%)

Available Information
A large body of intelligence on Fillistan's ballistic missile capacity has been gathered from multiple sources over time. This information is summarized in Table 1, which shows the main parameters for Fillistan's two known operational liquid-fuel missiles. For the purposes of this problem, you may assume that the information in Table 1 is accurate and has been independently corroborated. (credibility: almost certain 95-99%) (relevance: certain 100%)

Reason against hypothesis
A new source correctly reports that the missile was of Victory type
(force: likely 55-70%)

Sub-reason for
Direct evidence:
In late February 2017, the new source in Fillistan's General Staff reported that the missile tested on 10 January was a Victory missile. (credibility: likely 55-70%) (relevance: certain 100%)
(force: likely 55-70%)

Sub-reason against
Counter-intelligence got very active in Fillistan before the new source appeared
(force: barely likely 50-55%)

Evidence

In late December 2016, the source who works for the Best Flight R&D Institute was arrested, according to an intercept of an encrypted Fillistan Interior Ministry communication. (credibility: almost certain 95-99%) (relevance: more than likely 70-80%)

In early January 2017, the source in the guidance department at Fillistan's High Flight Guidance Institute detected Interior Ministry surveillance over a period of several days and requested immediate extraction from Fillistan. (credibility: certain 100%) (relevance: very likely 80-95%)

In late December 2016, a source who works in Fillistan's Missile Materials R&D Institute was summarily executed, according to an encrypted Interior Ministry communication. The source, according to the intercepted message, refused to provide any information during his interrogation. (credibility: almost certain 95-99%) (relevance: more than likely 70-80%)

Hypothesis 2: A missile of type Revolution with a modified inertial guidance system was launched : barely likely 50-55%
Figure for Paragraph 16

Reason for hypothesis
Of the two operational missile types Fillistan possesses, only Revolution can fly 500 km, so it could be Revolution
(force: barely likely 50-55%)
Figure for Paragraph 17

Sub-reason for
Known missile stats : almost certain 95-99%
Flew 500 km : almost certain 95-99%
(force: almost certain 95-99%)

Evidence

Available Information
A large body of intelligence on Fillistan's ballistic missile capacity has been gathered from multiple sources over time. This information is summarized in Table 1, which shows the main parameters for Fillistan's two known operational liquid-fuel missiles. For the purposes of this problem, you may assume that the information in Table 1 is accurate and has been independently corroborated. (credibility: almost certain 95-99%) (relevance: certain 100%)

On 10 January 2017, the Observation (a Buland missile-tracking radar ship deployed 150 km off Fillistan's coast detected a missile rising from Fillistan territory. The launch location was derived by back-tracking the trajectory and was assessed to be within a 1.5 km radius of the Matana launch complex. The missile's heading was due west, and flew directly over the ship. The missile impacted the ocean 500 km down-range from the launch point. All radar and tracking systems on the Observation were operational and recently calibrated. (credibility: almost certain 95-99%) (relevance: certain 100%)

Reason for hypothesis
Modifications were being made, and according to reported timelines were completed before the launch date, to the inertial guidance system of Revolution, and needed and were ready to be tested on January 10, 2017
(force: barely likely 50-55%)
Figure for Paragraph 21

Sub-reason for
Modifications : very likely 80-95%
Completion timeline : very likely 80-95%
Modifications can make a missile more range-flexible : more than likely 70-80%
Inertial guidance modifications need an actual launch to be tested : likely 55-70%
(force: likely 55-70%)

Sub-reason for
High command was interested in trajectory analysis
(force: likely 55-70%)

Evidence

In October 2016, a reliable source in the guidance department at the High Flight Guidance R&D Institute at Poolton reported that modifications were being made to the inertial guidance system (a guidance system that uses a computer and motion sensors) of the Revolution missile. (credibility: very likely 80-95%) (relevance: certain 100%)

In late November 2016, the head of the institute responsible for designing inertial guidance systems told Fillistan's Defense Minister that the design changes for the Revolution missile inertial guidance system would be completed by the end of the year. This exchange was captured in an intercept of an encrypted communication. (credibility: almost certain 95-99%) (relevance: very likely 80-95%)

According to an intercept of an encrypted message, on 1 February 2017, Jay Wilson (a senior Fillistan defense official in Umbabwea) called an individual identified as Alan Sampson at an unidentified R&D institute and asked whether he had completed analysis of the missile telemetry data from the 10 January test. (credibility: almost certain 95-99%) (relevance: more than likely 70-80%)

Assumption and Collection Requirement

Modifications can make a missile more range-flexible (credibility: more than likely 70-80%)

Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

Inertial guidance modifications need an actual launch to be tested (credibility: likely 55-70%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

Reason for hypothesis
Andrew Hall received the Hero award for work on the Revolution missile.
(force: barely likely 50-55%)
Figure for Paragraph 29

Sub-reason for
Hall received an award after the January 10, 2017 launch : certain 100%
The award was reportedly for the Revolution missile : likely 55-70%
(force: barely likely 50-55%)

Sub-reason against
Hall could not receive the award for the Revolution missile
(force: lacking support 0-50%)

Evidence

Andrew Hall was given a national "Hero" award in a highly publicized ceremony in early March 2017 for his accomplishments in "advancing Fillistan's national security." (credibility: certain 100%) (relevance: certain 100%)

A reliable source who works in Fillistan's Defense Ministry (the department responsible for all of Fillistan's missile development) reported in late March 2017 that Fillistan did not have any solid-fueled missiles in design or under development. The source also reported that Andrew Hall received the Hero award for his work on the Revolution missile. The source has reported reliably for ten years on a variety of defense issues. The source now works at a remote location and there had been no contact with the source for more than a year. In March 2017, the source travelled to the capital where contact was made. (credibility: likely 55-70%) (relevance: certain 100%)

In late December 2016, the source who works for the Best Flight R&D Institute was arrested, according to an intercept of an encrypted Fillistan Interior Ministry communication. (credibility: almost certain 95-99%) (relevance: more than likely 70-80%)

In early January 2017, the source in the guidance department at Fillistan's High Flight Guidance Institute detected Interior Ministry surveillance over a period of several days and requested immediate extraction from Fillistan. (credibility: certain 100%) (relevance: very likely 80-95%)

In late December 2016, a source who works in Fillistan's Missile Materials R&D Institute was summarily executed, according to an encrypted Interior Ministry communication. The source, according to the intercepted message, refused to provide any information during his interrogation. (credibility: almost certain 95-99%) (relevance: more than likely 70-80%)

Assumption and Collection Requirement

Hall could not be working on two different projects (credibility: likely 55-70%) (relevance: certain 100%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

Reason against hypothesis

There is a possibility that engineers only thought they were working on Revolution, while in fact the work was for a new missile because of compartmentalization
(force: lacking support 0-50%)

Assumption and Collection Requirement

There is a possibility that engineers only thought they were working on Revolution, while in fact the work was for a new missile because of compartmentalization (credibility: lacking support 0-50%) (relevance: barely likely 50-55%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

Hypothesis 3: A newly developed solid-fuel missile Progress was launched : lacking support 0-50%
Figure for Paragraph 40

Reason for hypothesis

Andrew Hall, the lead designer of Progress, received the Hero award for the project tested January 10, 2017
(force: likely 55-70%)
Figure for Paragraph 41

Sub-reason for

Andrew Hall, the lead designer of Progress, received the Hero award AFTER a project was tested on January 10, 2017 : certain 100%
Fillistan would reward the director of the institute responsible for the successful launch, as opposed to rewarding someone else to sow misinformation and protect the better scientist : likely 55-70%
(force: likely 55-70%)

Sub-reason against

Direct evidence:
A reliable source who works in Fillistan's Defense Ministry (the department responsible for all of Fillistan's missile development) reported in late March 2017 that Fillistan did not have any solid-fueled missiles in design or under development. The source also reported that Andrew Hall received the Hero award for his work on the Revolution missile. The source has reported reliably for ten years on a variety of defense issues. The source now works at a remote location and there had been no contact with the source for more than a year. In March 2017, the source travelled to the capital where contact was made. (credibility: barely likely 50-55%) (relevance: certain 100%)
(force: barely likely 50-55%)

Evidence

Andrew Hall was given a national "Hero" award in a highly publicized ceremony in early March 2017 for his accomplishments in "advancing Fillistan's national security." (credibility: certain 100%) (relevance: certain 100%)

Assumption and Collection Requirement

Fillistan would reward the director of the institute responsible for the successful launch, as opposed to rewarding someone else to sow misinformation and protect the better scientist (credibility: likely 55-70%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

Reason for hypothesis

It was not "Progress" that was launched in March : likely 55-70%
There were no other tests : likely 55-70%
(force: likely 55-70%)

Figure for Paragraph 46

Sub-reason for

Individuals behind the launched missile, director Doesky and deputy director Pullman, and the location of the institute at Pamplan, are different from lead Progress designer Andrew Hall at the Best Flight institute in Drogan
(force: likely 55-70%)
Figure for Paragraph 47

Sub-reason for

The March launch tested a project developed by the institute in Pamplan
(force: likely 55-70%)

Evidence

In June 2016, a reliable source in the Best Flight R&D Institute in the city of Drogan reported that Fillistan was designing a new solid-fuel missile, named Progress. At the time of the report, the Institute was about one year away from building and testing a prototype. The source began working for Best Flight after a transfer from another institute. (credibility: very likely 80-95%) (relevance: very likely 80-95%)

On 8 January 2017, an engineer at the Best Flight R&D Institute told a colleague that development of the Progress missile was going faster than expected, according to an intercept of an encrypted message. (credibility: almost certain 95-99%) (relevance: very likely 80-95%)

In June 2016, a reliable source in the Best Flight R&D Institute in the city of Drogan reported that Fillistan was designing a new solid-fuel missile, named Progress. At the time of the report, the Institute was about one year away from building and testing a prototype. The source began working for Best Flight after a transfer from another institute. (credibility: very likely 80-95%) (relevance: certain 100%)

According to an intercept of an encrypted message, on 22 December 2016, a Fillistan Defense Ministry official asked Andrew Hall (identified as the lead designer of the Progress missile) when the "solid-fueled missile" would be ready for testing. Hall responded, "in about three months." (credibility: almost certain 95-99%) (relevance: certain 100%)

According to an intercept of encrypted communications, on 1 March 2017, Alan Sampson called an individual identified as Jay Wilson and told him analysis of the telemetry data had been completed and was being reviewed by Mark Pullman, his institute's deputy director. Sampson noted that his institute was happy with the successful launch since all of the institute's efforts had gone into this crash program, the first missile this institute has been tasked with developing. The missile associated with the "crash program" was not identified. (credibility: almost certain 95-99%) (relevance: very likely 80-95%)

According to an intercept, the secretary of an unidentified R&D institute made airline reservations for an individual identified as Deputy Director Mark Pullman and another individual identified as Director John Doesky to fly from Pamplan to the capital of Fillistan on 30 April. (credibility: almost certain 95-99%) (relevance: certain 100%)

According to an intercept of encrypted communications, Marie Harris called John Doesky on 21 March 2017 and told him there is "nothing left of the missile," but the debris is being collected and will be sent to his institute at Pamplan. The institute was not identified. (credibility: almost certain 95-99%) (relevance: certain 100%)

According to an intercept, the secretary of an unidentified R&D institute made airline reservations for an individual identified as Deputy Director Mark Pullman and another individual identified as Director John Doesky to fly from Pamplan to the capital of Fillistan on 30 April. (credibility: almost certain 95-99%) (relevance: certain 100%)

According to an intercept of encrypted communications, Marie Harris called John Doesky on 21 March 2017 and told him there is "nothing left of the missile," but the debris is being collected and will be sent to his institute at Pamplan. The institute was not identified. (credibility: almost certain 95-99%) (relevance: certain 100%)

Assumption and Collection Requirement

The engineer at Best Flight knew the information because he works at Best Flight, and the reliable source at Best Flight knew about Progress because it was being developed at Best Flight (credibility: very likely 80-95%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

Flying from the same place and being mentioned together in a state with a highly compartmentalized research program means that they work at the same institute (credibility: likely 55-70%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

Director of an institute is the lead designer of the missile it is assigned developing, and design and assembly are not split between institutes (credibility: more than likely 70-80%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

Projects, even failed ones, do not get reassigned between institutions (credibility: likely 55-70%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

There were no other tests (credibility: likely 55-70%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

Reason against hypothesis
Progress was not complete on January 10, 2017
(force: likely 55-70%)
Figure for Paragraph 63

Sub-reason for
"Faster" was already accounted for in Hall's report
(force: likely 55-70%)

Sub-reason against
The development of Progress was going faster than expected : almost certain 95-99%
3 months more were needed before testing : very likely 80-95%
(force: barely likely 50-55%)

Evidence

In June 2016, a reliable source in the Best Flight R&D Institute in the city of Drogan reported that Fillistan was designing a new solid-fuel missile, named Progress. At the time of the report, the Institute was about one year away from building and testing a prototype. The source began working for Best Flight after a transfer from another institute. (credibility: very likely 80-95%) (relevance: more than likely 70-80%)

According to an intercept of an encrypted message, on 22 December 2016, a Fillistan Defense Ministry official asked Andrew Hall (identified as the lead designer of the Progress missile) when the "solid-fueled missile" would be ready for testing. Hall responded, "in about three months." (credibility: very likely 80-95%) (relevance: very likely 80-95%)

On 8 January 2017, an engineer at the Best Flight R&D Institute told a colleague that development of the Progress missile was going faster than expected, according to an intercept of an encrypted message. (credibility: almost certain 95-99%) (relevance: certain 100%)

According to an intercept of an encrypted message, on 22 December 2016, a Fillistan Defense Ministry official asked Andrew Hall (identified as the lead designer of the Progress missile) when the "solid-fueled missile" would be ready for testing. Hall responded, "in about three months." (credibility: very likely 80-95%) (relevance: very likely 80-95%)

Assumption and Collection Requirement

The program is compartmentalized - Hall was talking about his own research (credibility: very likely 80-95%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

Hypothesis 4: A newly developed solid-fuel missile, project name unknown, was launched : likely 55-70%
Figure for Paragraph 71

Reason for hypothesis
It could be an unknown new type of missile : likely 55-70%
It was not "Progress", but it was solid fuel : likely 55-70%
(force: likely 55-70%)
Figure for Paragraph 72

Sub-reason for
Funding and intelligence estimates of development costs suggest there might have been at least two solid-fuel missile projects in development in parallel as a way of hedging against research failures and intelligence leaks
(force: likely 55-70%)

Sub-reason for
Individuals behind the launched missile, director Doesky and deputy director Pullman, and the location of the institute at Pamplan, are different from lead Progress designer Andrew Hall at the Best Flight institute in Drogan : likely 55-70%
The missile launched on January 10, 2017 was a new solid fuel project : likely 55-70%
(force: likely 55-70%)

Sub-reason for
Progress was not complete on January 10, 2017
(force: likely 55-70%)

Evidence

According to an intercept of an encrypted communication, a Fillistan Defense Ministry official in January 2016 told another Defense Ministry official that development of solid-fuel missiles had become a high priority. Furthermore, he stated, funding for the development of solid-fuel missiles had increased by a factor of four, to a total of about $6 billion focused on swift creation of a functional prototype. (credibility: almost certain 95-99%) (relevance: certain 100%)

Buland's intelligence community estimates with 70 percent confidence that about $3 billion is needed for Fillistan to develop a new missile. (credibility: likely 55-70%) (relevance: certain 100%)

According to an intercept, the secretary of an unidentified R&D institute made airline reservations for an individual identified as Deputy Director Mark Pullman and another individual identified as Director John Doesky to fly from Pamplan to the capital of Fillistan on 30 April. (credibility: almost certain 95-99%) (relevance: certain 100%)

According to an intercept of encrypted communications, Marie Harris called John Doesky on 21 March 2017 and told him there is "nothing left of the missile," but the debris is being collected and will be sent to his institute at Pamplan. The institute was not identified. (credibility: almost certain 95-99%) (relevance: certain 100%)

According to an intercept of encrypted communications, on 1 March 2017, Alan Sampson called an individual identified as Jay Wilson and told him analysis of the telemetry data had been completed and was being reviewed by Mark Pullman, his institute's deputy director. Sampson noted that his institute was happy with the successful launch since all of the institute's efforts had gone into this crash program, the first missile this institute has been tasked with developing. The missile associated with the "crash program" was not identified. (credibility: almost certain 95-99%) (relevance: very likely 80-95%)

In June 2016, a reliable source in the Best Flight R&D Institute in the city of Drogan reported that Fillistan was designing a new solid-fuel missile, named Progress. At the time of the report, the Institute was about one year away from building and testing a prototype. The source began working for Best Flight after a transfer from another institute. (credibility: very likely 80-95%) (relevance: very likely 80-95%)

On 8 January 2017, an engineer at the Best Flight R&D Institute told a colleague that development of the Progress missile was going faster than expected, according to an intercept of an encrypted message. (credibility: almost certain 95-99%) (relevance: very likely 80-95%)

In June 2016, a reliable source in the Best Flight R&D Institute in the city of Drogan reported that Fillistan was designing a new solid-fuel missile, named Progress. At the time of the report, the Institute was about one year away from building and testing a prototype. The source began working for Best Flight after a transfer from another institute. (credibility: very likely 80-95%) (relevance: certain 100%)

According to an intercept of an encrypted message, on 22 December 2016, a Fillistan Defense Ministry official asked Andrew Hall (identified as the lead designer of the Progress missile) when the "solid-fueled missile" would be ready for testing. Hall responded, "in about three months." (credibility: almost certain 95-99%) (relevance: certain 100%)

According to an intercept of encrypted communications, on 1 March 2017, Alan Sampson called an individual identified as Jay Wilson and told him analysis of the telemetry data had been completed and was being reviewed by Mark Pullman, his institute's deputy director. Sampson noted that his institute was happy with the successful launch since all of the institute's efforts had gone into this crash program, the first missile this institute has been tasked with developing. The missile associated with the "crash program" was not identified. (credibility: almost certain 95-99%) (relevance: very likely 80-95%)

According to an intercept, the secretary of an unidentified R&D institute made airline reservations for an individual identified as Deputy Director Mark Pullman and another individual identified as Director John Doesky to fly from Pamplan to the capital of Fillistan on 30 April. (credibility: almost certain 95-99%) (relevance: certain 100%)

According to an intercept of encrypted communications, Marie Harris called John Doesky on 21 March 2017 and told him there is "nothing left of the missile," but the debris is being collected and will be sent to his institute at Pamplan. The institute was not identified. (credibility: almost certain 95-99%) (relevance: certain 100%)

According to an intercept, the secretary of an unidentified R&D institute made airline reservations for an individual identified as Deputy Director Mark Pullman and another individual identified as Director John Doesky to fly from Pamplan to the capital of Fillistan on 30 April. (credibility: almost certain 95-99%) (relevance: certain 100%)

According to an intercept of encrypted communications, on 1 March 2017, Alan Sampson called an individual identified as Jay Wilson and told him analysis of the telemetry data had been completed and was being reviewed by Mark Pullman, his institute's deputy director. Sampson noted that his institute was happy with the successful launch since all of the institute's efforts had gone into this crash program, the first missile this institute has been tasked with developing. The missile associated with the "crash program" was not identified. (credibility: almost certain 95-99%) (relevance: likely 55-70%)

In October 2016, a reliable source in the guidance department at the High Flight Guidance R&D Institute at Poolton reported that modifications were being made to the inertial guidance system (a guidance system that uses a computer and motion sensors) of the Revolution missile. (credibility: very likely 80-95%) (relevance: likely 55-70%)

In June 2016, a reliable source in the Best Flight R&D Institute in the city of Drogan reported that Fillistan was designing a new solid-fuel missile, named Progress. At the time of the report, the Institute was about one year away from building and testing a prototype. The source began working for Best Flight after a transfer from another institute. (credibility: very likely 80-95%) (relevance: more than likely 70-80%)

According to an intercept of an encrypted message, on 22 December 2016, a Fillistan Defense Ministry official asked Andrew Hall (identified as the lead designer of the Progress missile) when the "solid-fueled missile" would be ready for testing. Hall responded, "in about three months." (credibility: very likely 80-95%) (relevance: very likely 80-95%)

On 8 January 2017, an engineer at the Best Flight R&D Institute told a colleague that development of the Progress missile was going faster than expected, according to an intercept of an encrypted message. (credibility: almost certain 95-99%) (relevance: more than likely 70-80%)

According to an intercept of an encrypted message, on 22 December 2016, a Fillistan Defense Ministry official asked Andrew Hall (identified as the lead designer of the Progress missile) when the "solid-fueled missile" would be ready for testing. Hall responded, "in about three months." (credibility: very likely 80-95%) (relevance: very likely 80-95%)

Assumption and Collection Requirement

Flying from the same place and being mentioned together in a state with a highly compartmentalized research program means that they work at the same institute (credibility: likely 55-70%)
Deputy directors or directors were not promoted or transferred between institutes (credibility: likely 55-70%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

Projects, even failed ones, do not get reassigned between institutions (credibility: likely 55-70%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

The second launch can't fail if the first one was a success (credibility: likely 55-70%) (relevance: certain 100%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

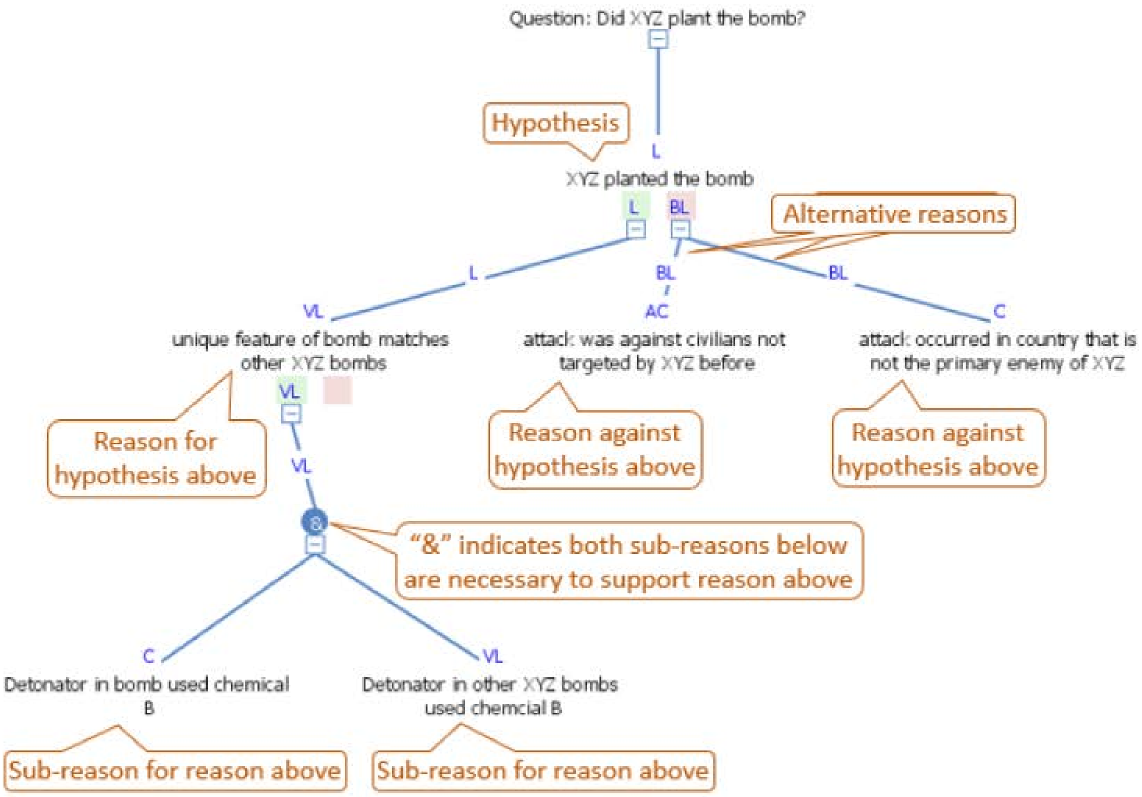
The engineer at Best Flight knew the information because he works at Best Flight, and the reliable source at Best Flight knew about Progress because it was being developed at Best Flight (credibility: very likely 80-95%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

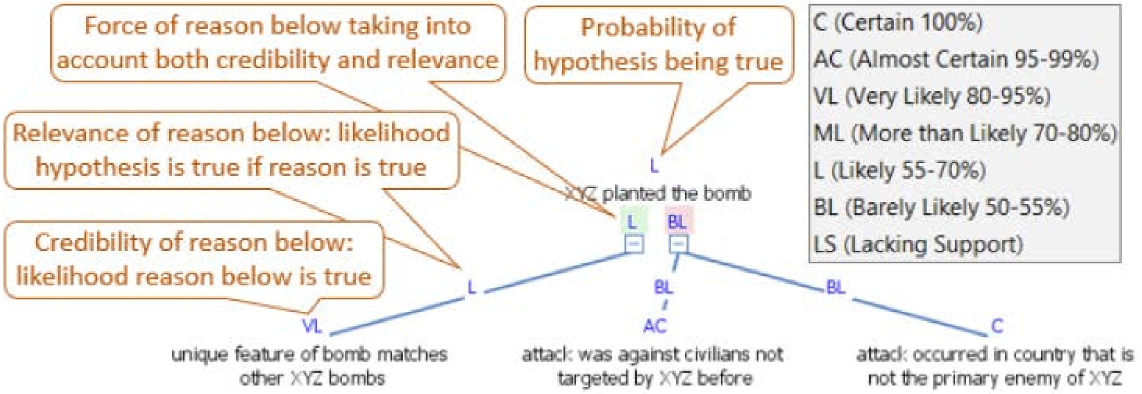
No other launches happened elsewhere during the time period (credibility: likely 55-70%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

Director of an institute is the lead designer of the missile it is assigned developing, and design and assembly are not split between institutes (credibility: more than likely 70-80%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

The program is compartmentalized - Hall was talking about his own research (credibility: very likely 80-95%)
Collect information that would either corroborate or contradict the validity of this assumption, to improve the analysis.

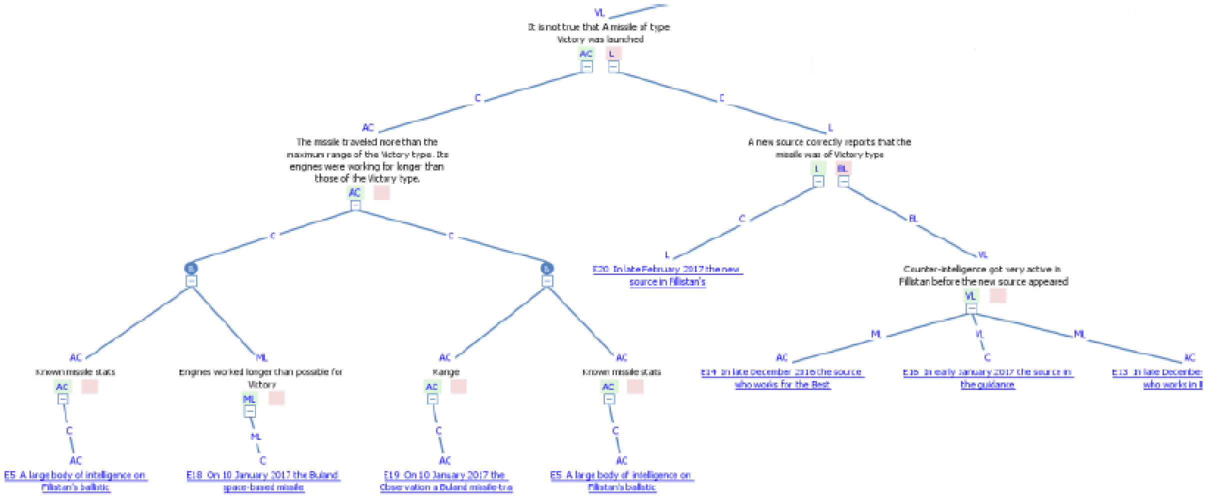
Review the Explanation of this Sample Analysis to Understand the Reasoning Diagrams in the Appendix


Question: Did XYZ plant the bomb?
Hypothesis
L
XYZ planted the bomb
L
BL
Alternative reasons
L
BL
AC
BL
C
VL
unique feature of bomb matches other XYZ bombs
attack was against civilians not targeted by XYZ before
attack occurred in country that is not the primary enemy of XYZ
VL
Reason for hypothesis above
Reason against hypothesis above
Reason against hypothesis above
VL
&
"&" indicates both sub-reasons below are necessary to support reason above
C
VL
Detonator in bomb used chemical B
Detonator in other XYZ bombs used chemcial B
Sub-reason for reason above
Sub-reason for reason above


Force of reason below taking into account both credibility and relevance
Probability of hypothesis being true
C (Certain 100%)
AC (Almost Certain 95-99%)
VL (Very Likely 80-95%)
ML (More than Likely 70-80%)
L (Likely 55-70%)
BL (Barely Likely 50-55%)
LS (Lacking Support)
Relevance of reason below: likelihood hypothesis is true if reason is true
L
XYZ planted the bomb
L
BL
Credibility of reason below: likelihood reason below is true
L
BL
BL
VL
AC
C
unique feature of bomb matches other XYZ bombs
attack was against civilians not targeted by XYZ before
attack occurred in country that is not the primary enemy of XYZ

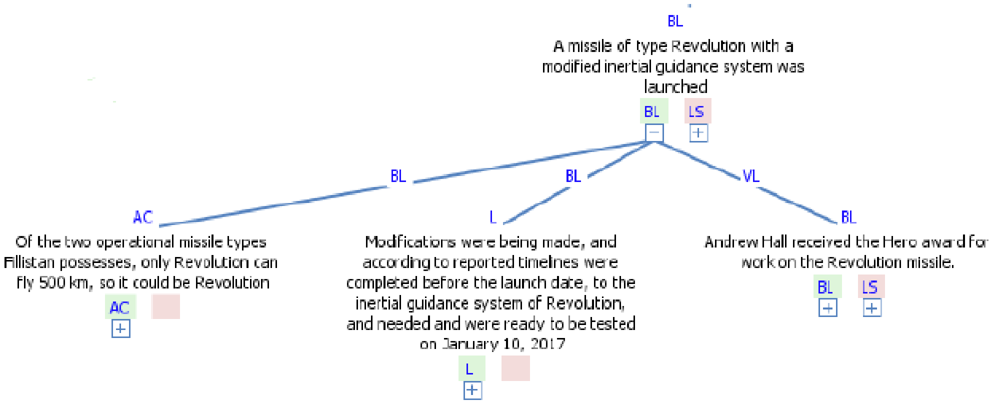
Appendix: What kind of ballistic missile was launched on 10 January 2017 from the Matana launch complex in Fillistan?

Figure for Paragraph 2:


VL
It is not true that A missile of type Victory was launched
AC
L
C
C
AC
The missile traveled more than the maximum range of the Victory type. Its engines were working for longer than those of the Victory type.
AC
L
A new source correctly reports that the missile was of Victory type
L
BL
C
C
C
BL
L
E20 In late February 2017 the new source in Fillistan's
VL
Counter-intelligence got very active in Fillistan before the new source appeared
VL
ML
VL
ML
AC
Known missile stats
AC
C
AC
E5 A large body of intelligence on Fillistan's ballistic
ML
Engines worked longer than possible for Victory
ML
ML
C
E18 On 10 January 2017 the Buland space-based missile
AC
Range
AC
C
AC
E19 On 10 January 2017 the Observation a Buland missile tra
AC
Known missile stats
AC
C
AC
E5 A large body of intelligence on Fillistan's ballistic
AC
E14 In late December 2016 the source who works for the Best
C
E15 In early January 2017 the source in the guidance
AC
E13 In late Decembe who works in I

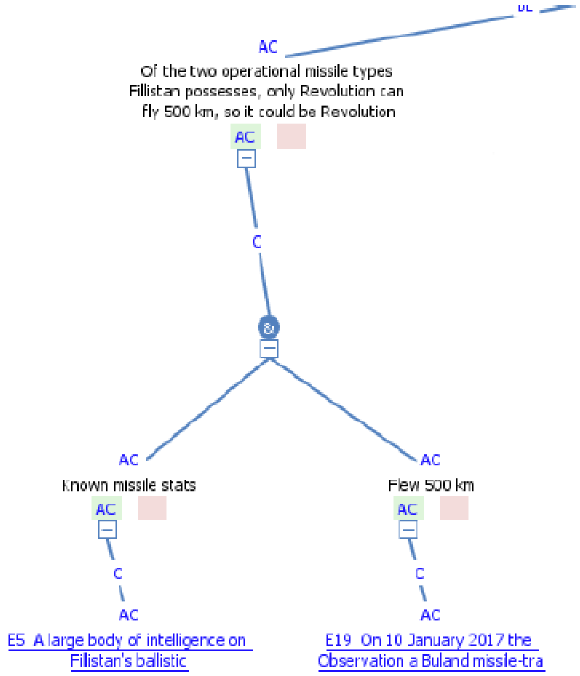
Return to Report

Figure for Paragraph 16:


BL
A missile of type Revolution with a modified inertial guidance system was launched
BL
LS
BL
BL
VL
AC
Of the two operational missile types Fillistan possesses, only Revolution can fly 500 km, so it could be Revolution
AC
L
Modifications were being made, and according to reported timelines were completed before the launch date, to the inertial guidance system of Revolution, and needed and were ready to be tested on January 10, 2017
L
BL
Andrew Hall received the Hero award for work on the Revolution missile.
BL
LS

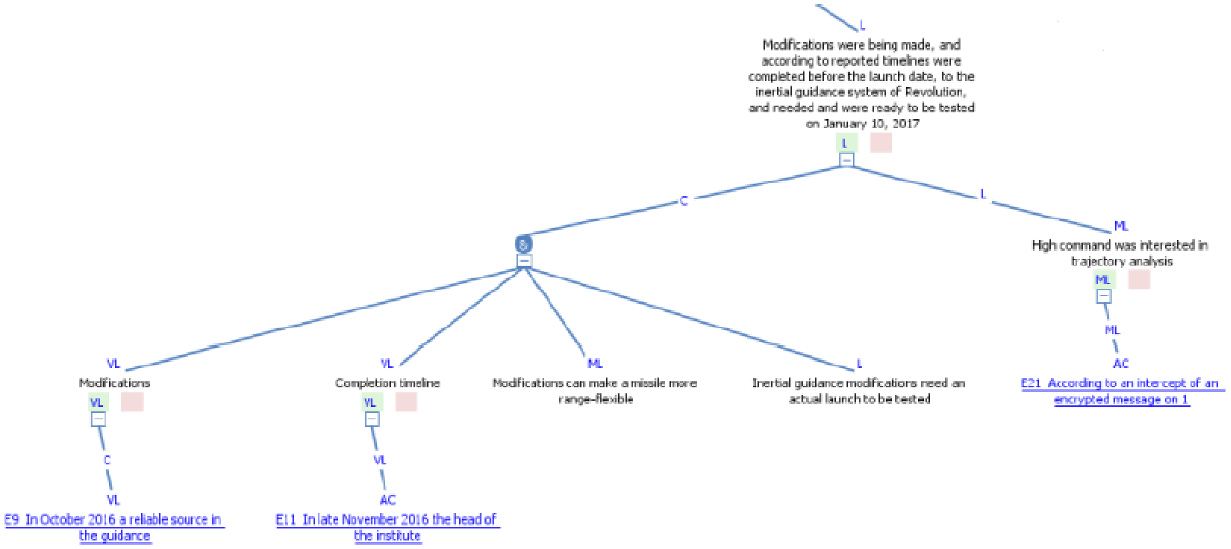
Return to Report

Figure for Paragraph 17:


AC
Of the two operational missile types Fillistan possesses, only Revolution can fly 500 km, so it could be Revolution
AC
C
&
AC
Known missile stats
AC
C
AC
E5 A large body of intelligence on Fillistan's ballistic
AC
Flew 500 km
AC
C
AC
E19 On 10 January 2017 the Observation a Buland missle-tra

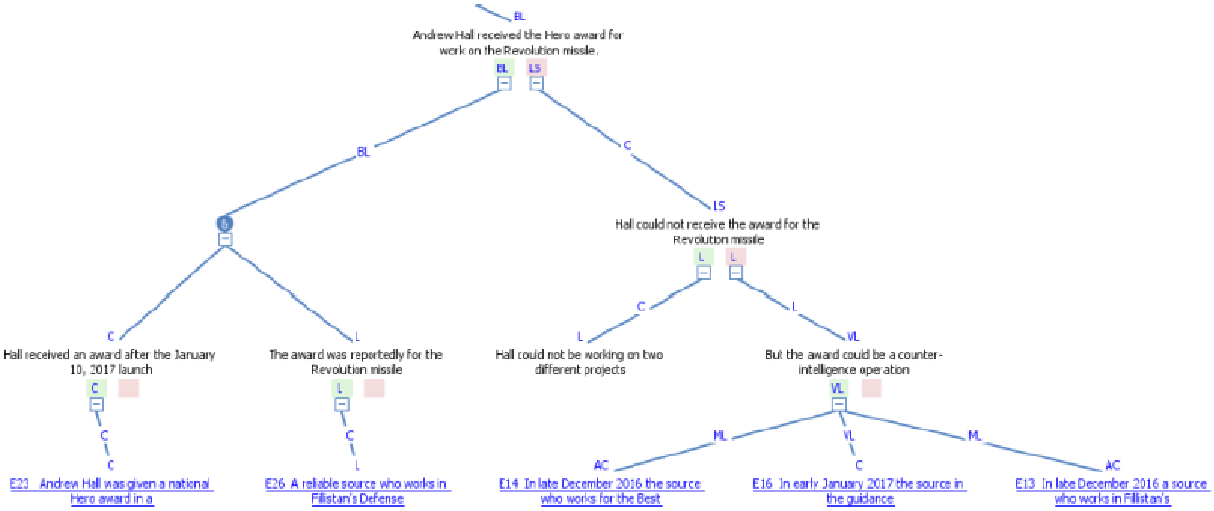
Return to Report

Figure for Paragraph 21:


L
Modifications were being made, and according to reported timelines were completed before the launch date, to the inertial guidance system of Revolution, and needed and were ready to be tested on January 10, 2017
L
C
L
&
VL
Modifications
VL
C
VL
E9 In October 2016 a reliable source in the guidance
VL
Completion timeline
VL
VL
AC
E11 In late November 2016 the head of the institute
ML
Modifications can make a missile more range-flexible
L
Inertial guidance modifications need an actual launch to be tested
ML
Hgh command was interested in trajectory analysis
ML
ML
AC
E21 According to an intercept of an encrypted message on 1

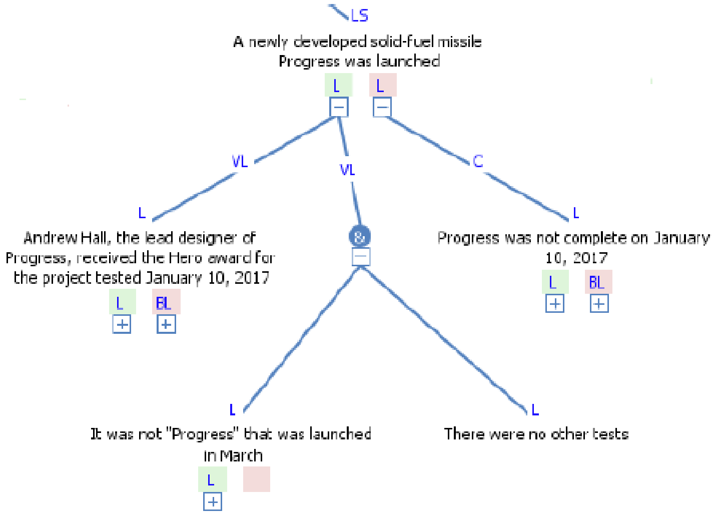
Return to Report

Figure for Paragraph 29:


BL
Andrew Hall received the Hero award for work on the Revolution missle.
BL
LS
BL
C
LS
&
Hall could not receive the award for the Revolution missile
L
L
C
Hall received an award after the January 10, 2017 launch
C
C
C
L
The award was reportedly for the Revolution missile
L
C
L
C
L
Hall could not be working on two different projects
L
VL
But the award could be a counter-intelligence operation
VL
ML
VL
ML
AC
C
AC
E23 Andrew Hall was given a national Hero award in a
E26 A reliable source who works in Fillstan's Defense
E14 In late December 2016 the source who works for the Best
E16 In early January 2017 the source in the guidance
E13 In late December 2016 a source who works in Fillstan's

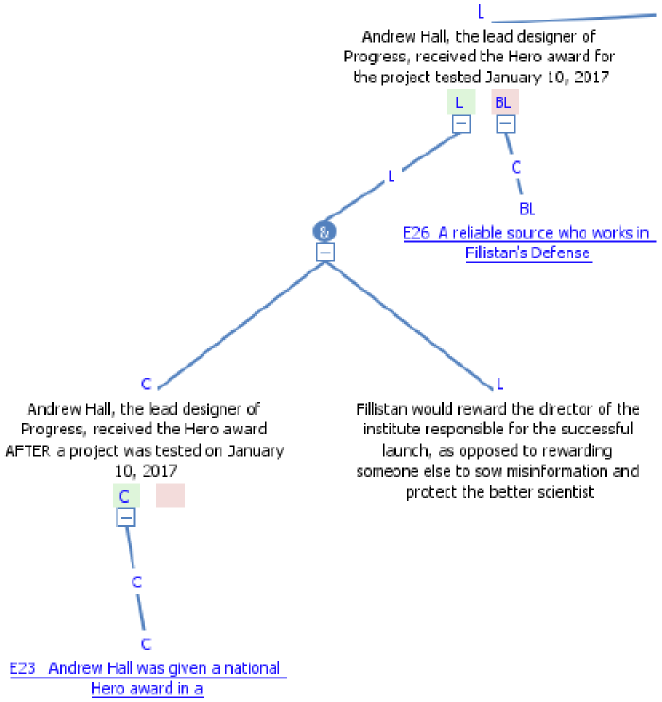
[Return to Report]

Figure for Paragraph 40:


LS
A newly developed solid-fuel missile Progress was launched
L
L
VL
VL
C
L
Andrew Hall, the lead designer of Progress, received the Hero award for the project tested January 10, 2017
L
BL
&
L
Progress was not complete on January 10, 2017
L
BL
L
It was not "Progress" that was launched in March
L
L
There were no other tests

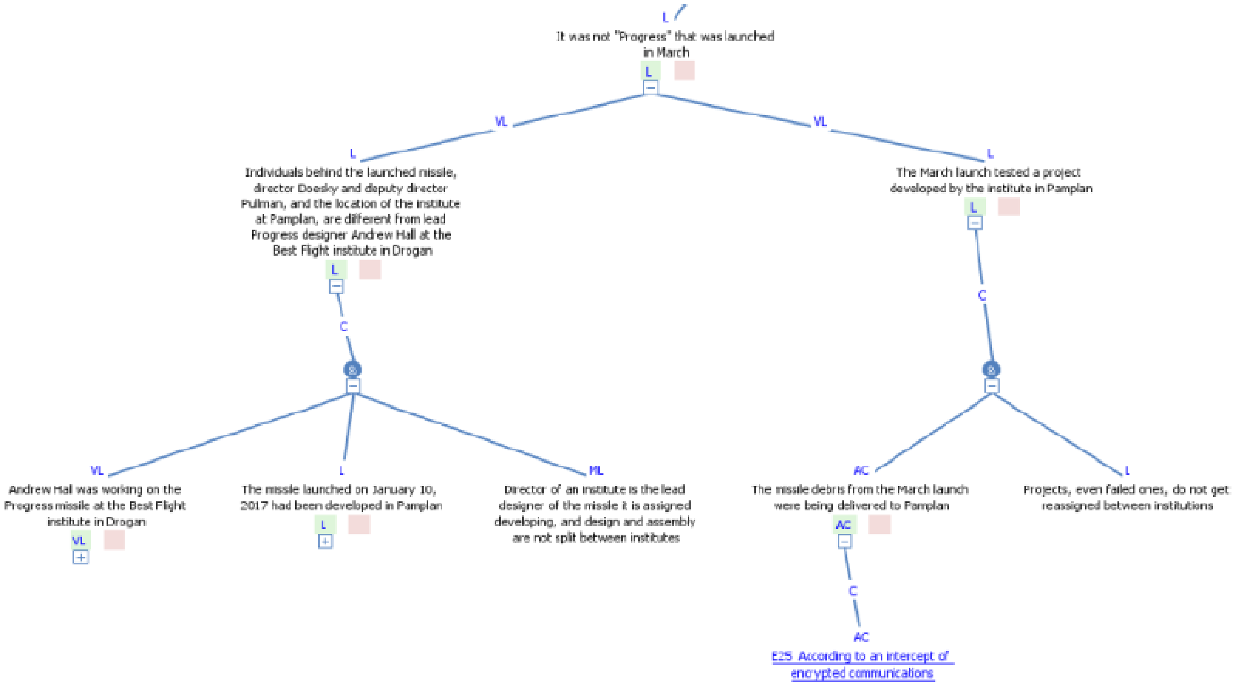
[Return to Report]

Figure for Paragraph 41:


L
Andrew Hall, the lead designer of Progress, received the Hero award for the project tested January 10, 2017
L
BL
L
C
BL
E26 A reliable source who works in Filistan's Defense
&
C
Andrew Hall, the lead designer of Progress, received the Hero award AFTER a project was tested on January 10, 2017
C
C
C
E23 Andrew Hall was given a national Hero award in a
L
Fillistan would reward the director of the institute responsible for the successful launch, as opposed to rewarding someone else to sow misinformation and protect the better scientist

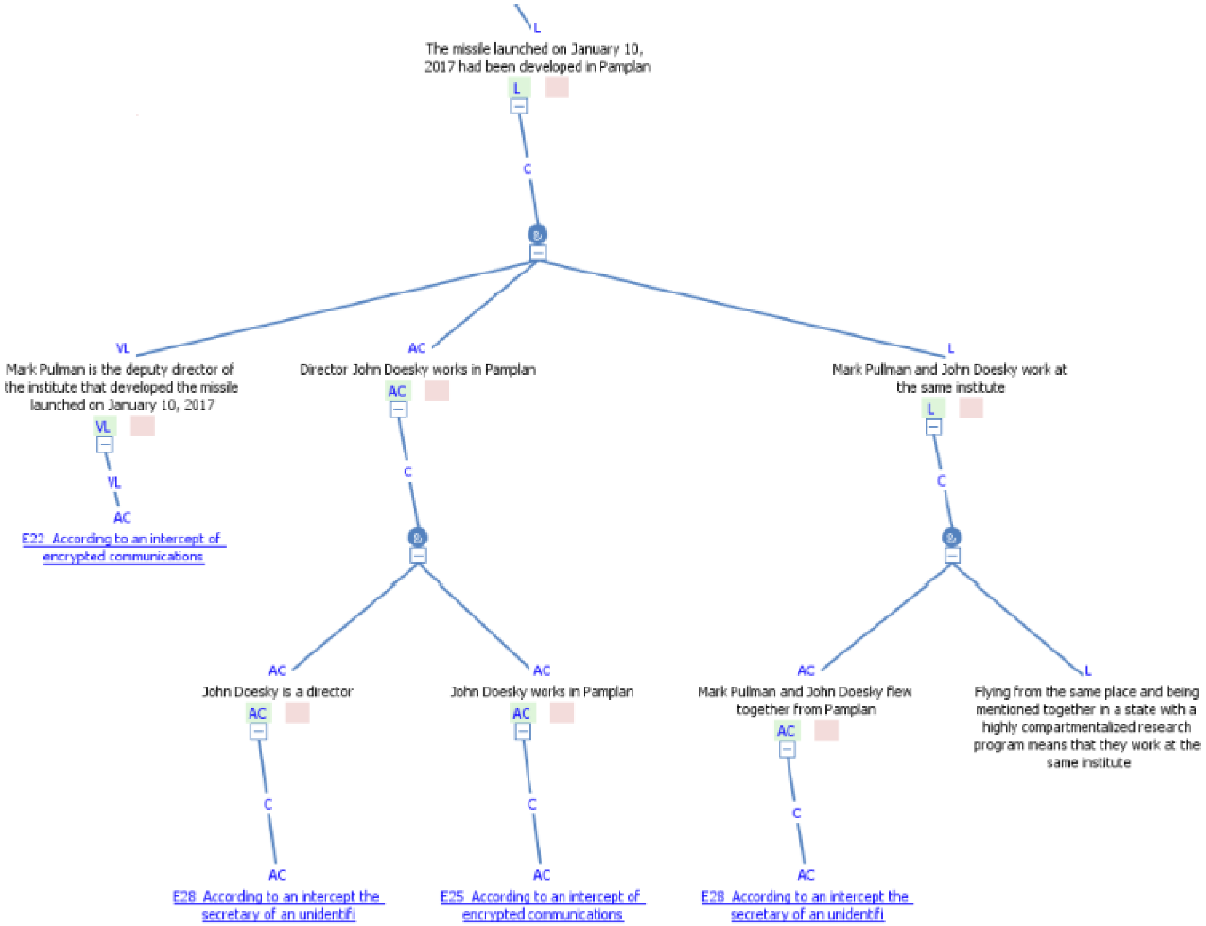
Return to Report

Figure for Paragraph 46:


L
It was not "Progress" that was launched in March
L
VL
L
Individuals behind the launched nissle, director Doesky and deputy director Pulman, and the location of the institute at Pamplan, are different from lead Progress designer Andrew Hall at the Best Flight institute in Drogan
L
C
&
VL
Andrew Hal was working on the Progress missile at the Best Flight institute in Drogan
VL
L
The missle launched on January 10, 2017 had been developed in Pamplan
L
ML
Director of an institute is the lead designer of the missle it is assigned developing, and design and assembly are not split between institutes
VL
L
The March launch tested a project developed by the institute in Pamplan
L
C
&
AC
The missle debris from the March launch were being delivered to Pamplan
AC
C
AC
E25 According to an intercept of encrypted communications
L
Projects, even failed ones, do not get reassigned between institutions

Return to Report

Figure for Paragraph 47:



The missile launched on January 10, 2017 had been developed in Pamplan

Mark Pulman is the deputy director of the institute that developed the missile launched on January 10, 2017

Director John Doesky works in Pamplan

Mark Pullman and John Doesky work at the same institute

E22 According to an intercept of encrypted communications

John Doesky is a director

John Doesky works in Pamplan

Mark Pullman and John Doesky flew together from Pamplan

Flying from the same place and being mentioned together in a state with a highly compartmentalized research program means that they work at the same institute

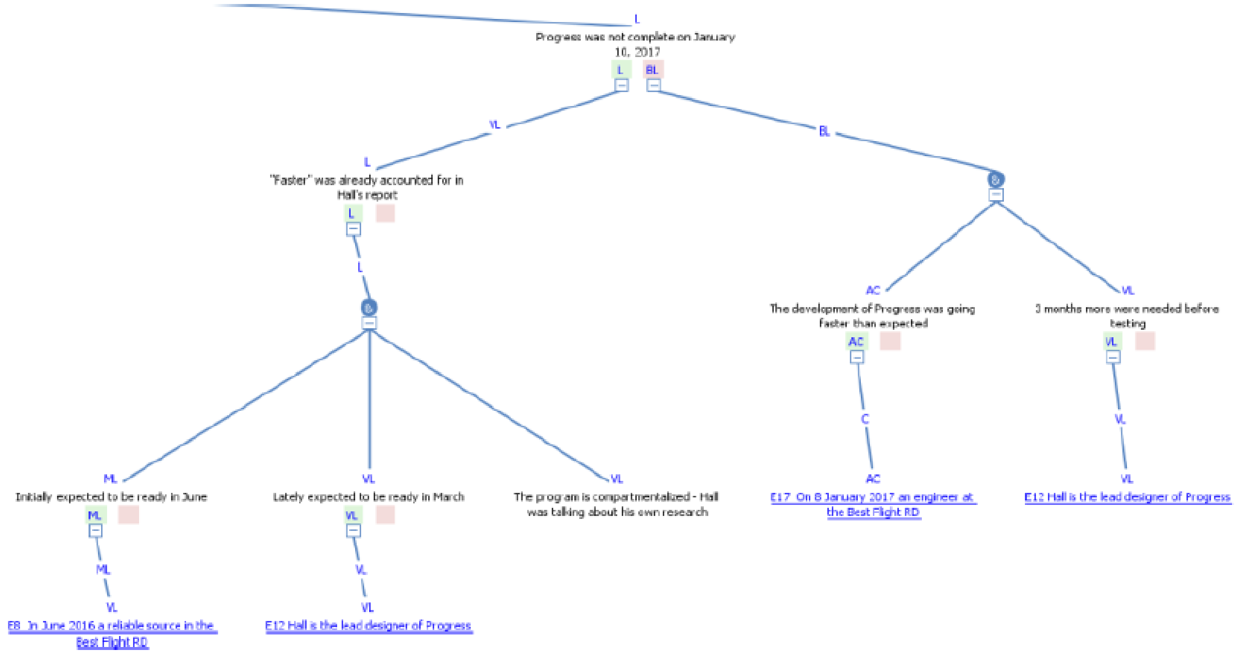
E28 According to an intercept the secretary of an unidentifi

E25 According to an intercept of encrypted communications

E28 According to an intercept the secretary of an unidentifi

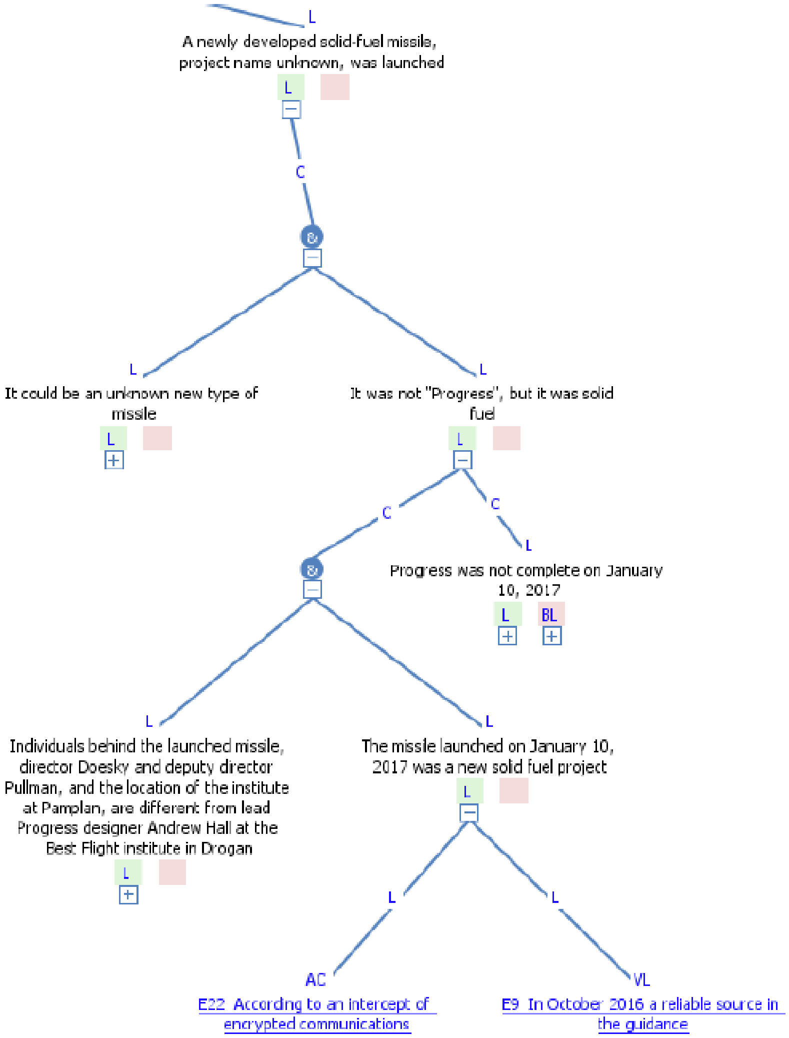
[Return to Report]

Figure for Paragraph 63:


L
Progress was not complete on January 10, 2017
L
BL
VL
BL
L
"Faster" was already accounted for in Hall's report
L
L
&
ML
Initially expected to be ready in June
ML
ML
VL
E8 In June 2016 a reliable source in the Best Flight RD
VL
Lately expected to be ready in March
VL
VL
VL
E12 Hall is the lead designer of Progress
VL
The program is compartmentalized - Hall was talking about his own research
&
AC
The development of Progress was going faster than expected
AC
C
AC
E17 On 8 January 2017 an engineer at the Best Flight RD
VL
3 months more were needed before testing
VL
VL
VL
E12 Hall is the lead designer of Progress

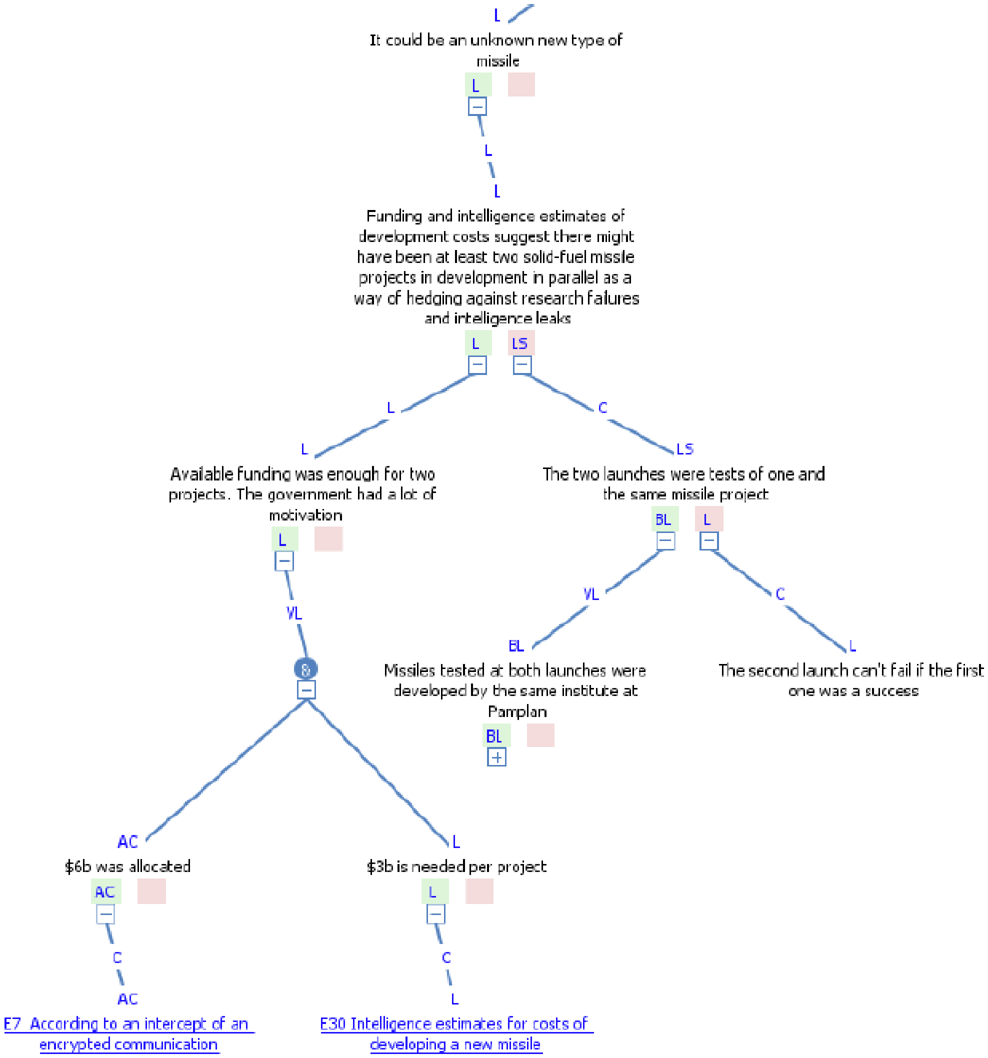
Return to Report

Figure for Paragraph 71:


L
A newly developed solid-fuel missile, project name unknown, was launched
L
C
&
L
It could be an unknown new type of missile
L
L
It was not "Progress", but it was solid fuel
L
C
C
&
L
Progress was not complete on January 10, 2017
L
BL
L
Individuals behind the launched missle, director Doesky and deputy director Pullman, and the location of the institute at Pamplan, are different from lead Progress designer Andrew Hall at the Best Flight institute in Drogan
L
L
The missle launched on January 10, 2017 was a new solid fuel project
L
L
L
AC
E22 According to an intercept of encrypted communications
VL
E9 In October 2016 a reliable source in the guidance

Return to Report

Figure for Paragraph 72:


L
It could be an unknown new type of missile
L
L
L
Funding and intelligence estimates of development costs suggest there might have been at least two solid-fuel missle projects in development in parallel as a way of hedging against research failures and intelligence leaks
L
LS
L
L
C
LS
Available funding was enough for two projects. The government had a lot of motivation
The two launches were tests of one and the same missile project
L
BL
L
VL
&
VL
BL
C
L
Missiles tested at both launches were developed by the same institute at Pamplan
The second launch can't fail if the first one was a success
BL
AC
L
$6b was allocated
$3b is needed per project
AC
L
C
C
AC
L
E7 According to an intercept of an encrypted communication
E30 Intelligence estimates for costs of developing a new missile

Return to Report

9.9. “Who Is the Spy?” Problem

This is a fictional case, not based on real-world events or entities. All the information you need to answer this question is provided along with the problem. The information provided may not be entirely conclusive or comprehensive, and answers may not be absolutely certain. Please do not search for additional information online-- this is a test of reasoning rather than your ability to search for new clues. You should use your best judgment and reasoning in arriving at your answer, explain what alternatives you considered, and give your arguments for why a reader should believe your conclusions.

Who is the Spy?

Scenario

Situation: Three Arborian army intelligence officers-Captain Harry, Major Tom, and Lt. Col. Dick-work in the cryptology office C/M-2, whose primary responsibility is breaking the military communications codes of various countries, including Razmania and Plumistan. C/M-2 receives a variety of finished intelligence reports on a wide range of subjects. They work in Room 3C12, a highly secure, vaulted area to which only they are allowed unrestricted access, in Arboria's Central Cryptology Center (CCC), which occupies an enormous building in Monaca, Arboria's capital. Captain Harry, Major Tom, and Lt. Col. Dick are mathematicians.

An uncleared custodian named Paul Jones enters 3C12 every workday to empty the trash cans that are only authorized for unclassified material and rubbish. Because there are mounds of classified material everywhere, Harry, Tom, or Dick always escort Paul when he is in the vault and monitor his movements closely. They are careful not to discuss classified material when Paul is in the vault. At the end of the day, the vault is physically locked and an electronic security system is activated to ensure that there is no unregistered entry. No one other than Paul has partial access to 3C12.

Arboria and Razmania are allies of sorts, but not all of their interests coincide. As a result, Arboria does not share all of its intelligence with Razmania. Thus, Razmania's Intelligence Service (MIS) actively tries to recruit employees of Arboria's CCC as spies. Unbeknownst to Razmania, Arboria is able to break the diplomatic encryption codes used in Razmania's electronic communications.

Razmania's main threat is Plumistan, a large country that borders it to the north. Plumistan has publicly acknowledged that its main strategic goal is the annihilation of Razmania. Plumistan has also threatened to support terrorist attacks against Arboria.

Security logs show that on 1 November 2016, C/M-2 took receipt of copy no. 87 of the document "Military Encryption Codes of Plumistan," which contained information on the various cryptographic systems that Plumistan's military uses. Exactly 200 copies of this document were circulated in total.

In early January 2017, a reliable source working in the office of Razmania's Foreign Minister provided a copy of a document that had been sent from the head of Razmania's Intelligence Service (MIS) to Razmania's foreign and defense ministers:

At around noon on 9 December 2016, an individual identifying themselves as "Archimedes" tossed a copy of an Arborian intelligence document over the wall of our embassy in Monaca. The document was entitled "Military Encryption Codes of Plumistan" and the document was identified as No. 87 out of 200. Archimedes said they were a friend of Razmania, and the document enclosed was meant to show what kinds of information they could provide.

At around noon on 12 December 2016, the same individual tossed a copy of an Arborian intelligence document on Plumistan's new tank over the wall of our embassy in Monaca. Archimedes asked that this information not be communicated back to Razmania via electronic communications but be hand carried back (in a diplomatic pouch that cannot be opened or otherwise searched) to Razmania. Taped to this document was a message that again noted Archimedes' concern about Razmania's security but also offered long-term cooperation with Razmania in exchange for money.

At around 10 AM on 18 December 2016, "Archimedes" tossed a message over the wall of our embassy in Monaca. This message included instructions for leaving $25,000 at a dead drop site along with additional instructions for receiving additional documents at another dead drop site. The message again warned against communicating this information via electronic communication systems.

Arboria's videotape of the individual tossing the documents over the wall is not clear. The weather on all three days when the documents were tossed over the wall of Razmania's Embassy was very cold and windy, and the individual was wearing sunglasses and a heavy, hooded coat. Razmania's Embassy typically is about a 30-minute drive from CCC's headquarters in Monaca.

You are a counterintelligence analyst working for Arboria.

Key Question

Is Tom, Dick, Harry, or Paul the person removing the classified information that has been passed to Razmania from C/M-2?

Available Information

Note: Assume all biographic information prior to June 2016 is true. Also assume that there is no collaboration between Dick, Tom, Harry, and Paul.

- Security logs show that the 3C12 vault has been locked and properly secured at the end of every workday for the past year.

- Captain Harry is single and has never been married. He lives alone and has no close friends. People who have worked with Harry believe he is a bit of a recluse. Some of this is due to Harry having Asperger's Syndrome, and some of this is due to Harry's experience growing up. Harry's parents were killed in a car accident when Harry was 18 months old, and he moved from foster home to foster home during the first ten years of his life. A 2010 psychological assessment of Harry (conducted when Harry was commissioned as an army officer) noted that Harry has "trust"" issues.

- Harry is a math and computer-programming wizard, considered brilliant with an IQ over 165. Many of his civilian co-workers refer to Harry as "the professor." Some also call him the "absent-minded professor" because Harry can be forgetful, especially when he is engaged in a major intellectual undertaking. Harry spends almost every night of the week at home in his apartment, participating in online forums about the history of mathematics. According to neighbors, Harry receives no in-person visitors-his social life is almost exclusively online.

- Harry's mother was born and raised in Razmania. In office conversations, Harry is quite vocal about Razmania's geopolitical vulnerability and is adamant that Arboria is not doing enough to defend Razmania. Most people in Harry's office no longer talk about Razmania in Harry's presence because of Harry's emotional outbursts during these discussions. Harry has traveled to Razmania at least once a year for the past seven years. He has reported all of these trips to the Office of Security in C/M-2.

- Major Tom is married. When Tom married Betsy, people said Tom married "up." Betsy grew up in a very wealthy family and her friends say that Betsy is addicted to having "all the finer things in life." She is upset that her sisters live in bigger houses and drive high-end SUVs. Tom, well aware that he is not providing Betsy with everything she wants, is always trying to please her. Tom is what people called a "straight arrow"-always following the rules.

- Lt. Colonel Dick, who is the head of C/M-2, is married to Sally, a beautiful woman 15 years younger. They have no children, although he has two children from a previous marriage. She is a "little rough around the edges," according to Dick's friends. She was arrested for shoplifting as a teenager. Dick's younger sister Molly Parker has remarked that Sally has no morals and is a "natural criminal." The relationship between Sally and Molly turned ugly two years ago (in 2015) immediately after Sally accused Molly's husband John Parker of making a pass at her. Molly responded by calling Sally a "no-good liar." During his two-year marriage to Molly, John has had three affairs.

- Paul, age 45, has had a rough life. His father was abusive and his mother was a cocaine addict. He dropped out of high school in the 10th grade. As a young man in his early twenties, Paul committed a number of felonies, including armed robbery, according to state penal records. Paul spent four years in prison. While in prison, Paul found God and turned his life around. He has not had a speeding ticket in 20 years. When not working, Paul spends most of his day watching reality TV. A friend of Paul's once said that Paul probably could not name more than three capitals in the world, not counting Monaca.

- Paul has an older daughter, named Jennifer, who he adores. Jennifer needs a liver transplant, but does not have insurance. The cost for the transplant is about $100,000. Paul is not sleeping well because of his daughter's condition.

- Dick loves to gamble. Every year Dick and his wife travel to Palisade, Arboria's gambling mecca in the mountains. Investigators discovered that in June 2016, Dick lost $30,000, and in October 2016, he hired a financial adviser who recommended that Dick take out a second-equity loan on his house, which Dick did. Relatives of Dick and Sally told investigators that Dick and Sally argue a lot about money.

- On 10 July 2016, Harry had dinner at the Great Inn in downtown Monaca with Dr. Mark Down, who is listed as a Cultural Attache in the Razmanian Embassy. According to CCC security records, Harry reported this contact. Arboria's internal security agency suspects that Down's real employer is the MIS and that his cultural-attache title is just a cover. However, this is not known with certainty. In his report, Harry noted that he met Down during one of his visits to Razmania, and they both share a deep interest in Razmania's history.

- In September 2016, military records indicate that Tom was passed up for promotion for the second time. One of Dick's comments in Tom's personnel file was that Tom "did not live up to his potential and therefore should not be promoted."

- According to investigators, Tom's and Betsy's best friends, Bill and Julie, said that Tom blamed Lt. Col. Dick for not getting promoted and said Dick "screwed him" and "the Army can go to Hell." Tom said that this reminded him of the first time he failed to be promoted because his commanding officer did not support his promotion. Bill said Tom was livid. After being passed up for promotion a second time, Tom, according to Army regulations, cannot be promoted and must retire within two years. Julie said that Betsy was even angrier about Tom not getting promoted than Tom was. She wanted to buy a bigger house. The available credit on four of their five credits cards was almost zero in July 2016, according to credit card records.

- In October 2016, Dick was passed up for promotion to full colonel. The promotion would not have made Dick rich, but the extra money and perks would have eased some of the stress in making his loan and credit card obligations, allowing for more discretionary spending. Dick told Greg Holland, his best friend, that he was confident he would get promoted in the next promotion cycle.

- Despite adultery being a criminal offense under Arboria's Uniform Code of Military Justice, Dick began an affair with Becky, a young woman, in early November 2016. According to investigators, Dick only told Greg Holland about the affair. Dick told Greg in late November that he likes Becky but he doesn't see the relationship going anywhere. He also told Greg that he "hated" the everyday pressure that Sally put on him regarding finances and that he was going to seek a divorce. Dick also said that Sally was one of the most superficial people he had ever met, and he doesn't know why he ever asked her to marry him.

- In November 2016, Tom and Betsy travelled to Tralia for a week-long vacation, according to airline records. However, Tom did not submit a foreign travel request to CCC security, as required. Tom charged all the expenses for this trip, according to credit card records.

- Outside of Greg, Dick has no real friends. Greg, who is retired, travelled to Tralia and stayed there for the entire month of December, according to travel records. Greg invited Dick to join him in Tralia, but Dick said he couldn't afford the airfare and was too busy at work.

- Arboria's internal security agency reviewed the credit card records of Harry and Dr. Mark Down in February 2017, and discovered that both Harry and Down paid bills at the Great Inn on 15 October 2016 within several minutes of each other. Harry did not report any contact with Down on or near that date, according to CCC security records. At this time, it is uncertain whether such contact occurred.

- In late September 2016, Plumistan began to use a new encryption system for its diplomatic messages. Harry was temporarily assigned to the team that was tasked with breaking this new encryption system. The system was broken in early November after team members-including Harry-worked several 65-hour-work weeks running computer program after computer program.

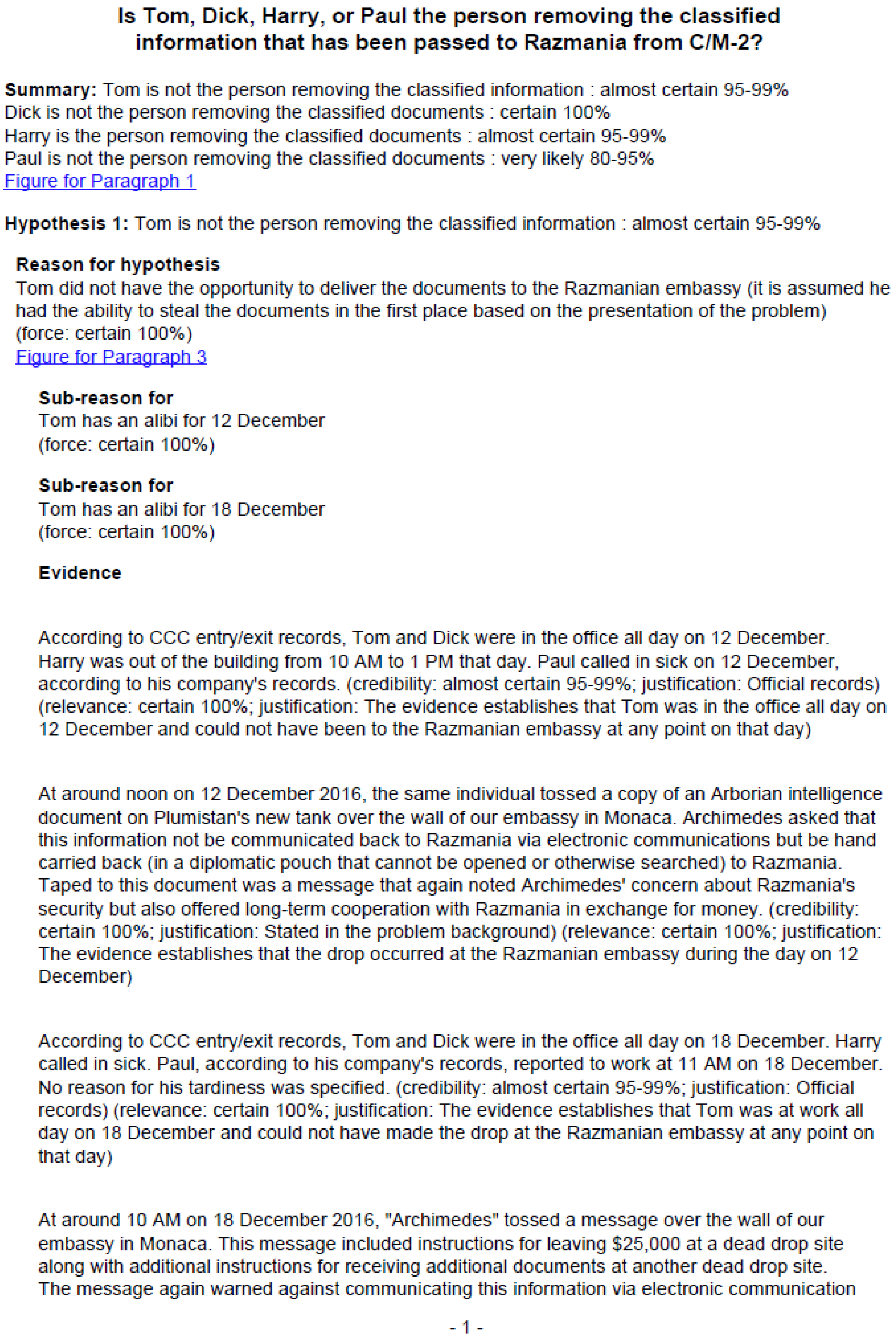
- Dick is 50 years old. In early December 2016, he had a medical physical. Dick told his best friend Greg that the doctor read him the "riot act" about his blood pressure and that he needed to quit smoking and drinking immediately. To the surprise of almost everyone, including his doctor, Dick quit smoking and drinking. Dick has not resumed smoking and drinking, to date.

- According to CCC entry/exit records, Tom and Harry were in the office at noon on 9 December 2016. Entry/exit records show that Dick was out of the office from 11AM to 1PM on this day. Paul Jones was in CCC's headquarters from 8AM to 11AM on 9 December, according to entry and exit logs for CCC's headquarters.

- According to CCC entry/exit records, Tom and Dick were in the office all day on 12 December. Harry was out of the building from 10 AM to 1 PM that day. Paul called in sick on 12 December, according to his company's records.

- According to CCC entry/exit records, Tom and Dick were in the office all day on 18 December. Harry called in sick. Paul, according to his company's records, reported to work at 11 AM on 18 December. No reason for his tardiness was specified.

9.10. The Solution “aaron83” of the “Spy” Problem in the T&E Experiment



Is Tom, Dick, Harry, or Paul the person removing the classified information that has been passed to Razmania from C/M-2?

Summary: Tom is not the person removing the classified information : almost certain 95-99%
Dick is not the person removing the classified documents : certain 100%
Harry is the person removing the classified documents : almost certain 95-99%
Paul is not the person removing the classified documents : very likely 80-95%
Figure for Paragraph 1

Hypothesis 1: Tom is not the person removing the classified information : almost certain 95-99%

Reason for hypothesis
Tom did not have the opportunity to deliver the documents to the Razmanian embassy (it is assumed he had the ability to steal the documents in the first place based on the presentation of the problem)
(force: certain 100%)
Figure for Paragraph 3

Sub-reason for
Tom has an alibi for 12 December
(force: certain 100%)

Sub-reason for
Tom has an alibi for 18 December
(force: certain 100%)

Evidence

According to CCC entry/exit records, Tom and Dick were in the office all day on 12 December. Harry was out of the building from 10 AM to 1 PM that day. Paul called in sick on 12 December, according to his company's records. (credibility: almost certain 95-99%; justification: Official records) (relevance: certain 100%; justification: The evidence establishes that Tom was in the office all day on 12 December and could not have been to the Razmanian embassy at any point on that day)

At around noon on 12 December 2016, the same individual tossed a copy of an Arborian intelligence document on Plumistan's new tank over the wall of our embassy in Monaca. Archimedes asked that this information not be communicated back to Razmania via electronic communications but be hand carried back (in a diplomatic pouch that cannot be opened or otherwise searched) to Razmania. Taped to this document was a message that again noted Archimedes' concern about Razmania's security but also offered long-term cooperation with Razmania in exchange for money. (credibility: certain 100%; justification: Stated in the problem background) (relevance: certain 100%; justification: The evidence establishes that the drop occurred at the Razmanian embassy during the day on 12 December)

According to CCC entry/exit records, Tom and Dick were in the office all day on 18 December. Harry called in sick. Paul, according to his company's records, reported to work at 11 AM on 18 December. No reason for his tardiness was specified. (credibility: almost certain 95-99%; justification: Official records) (relevance: certain 100%; justification: The evidence establishes that Tom was at work all day on 18 December and could not have made the drop at the Razmanian embassy at any point on that day)

At around 10 AM on 18 December 2016, "Archimedes" tossed a message over the wall of our embassy in Monaca. This message included instructions for leaving $25,000 at a dead drop site along with additional instructions for receiving additional documents at another dead drop site. The message again warned against communicating this information via electronic communication systems. (credibility: certain 100%; justification: Stated in the problem background) (relevance: certain 100%; justification: The evidence establishes that the drop at the Razmanian embassy occurred during the day on 18 December.)

Reason against hypothesis
Tom had the means
(force: barely likely 50-55%)
Figure for Paragraph 10

Sub-reason for
Tom does not always follow the rules
(force: barely likely 50-55%)

Evidence

In November 2016, Tom and Betsy travelled to Tralia for a week-long vacation, according to airline records. However, Tom did not submit a foreign travel request to CCC security, as required. Tom charged all the expenses for this trip, according to credit card records. (credibility: almost certain 95-99%; justification: Based on CCC observation and records) (relevance: certain 100%; justification: The evidence shows that Tom was willing to break the rules in order to make his wife happy)

Major Tom is married. When Tom married Betsy, people said Tom married "up." Betsy grew up in a very wealthy family and her friends say that Betsy is addicted to having "all the finer things in life." She is upset that her sisters live in bigger houses and drive high-end SUVs. Tom, well aware that he is not providing Betsy with everything she wants, is always trying to please her. Tom is what people called a "straight arrow"-always following the rules. (credibility: certain 100%; justification: Stated by the problem) (relevance: likely 55-70%; justification: The evidence shows that Tom has a reputation for probity, regardless of whether it is true or not)

Reason against hypothesis
Tom had the motive
(force: barely likely 50-55%)
Figure for Paragraph 14

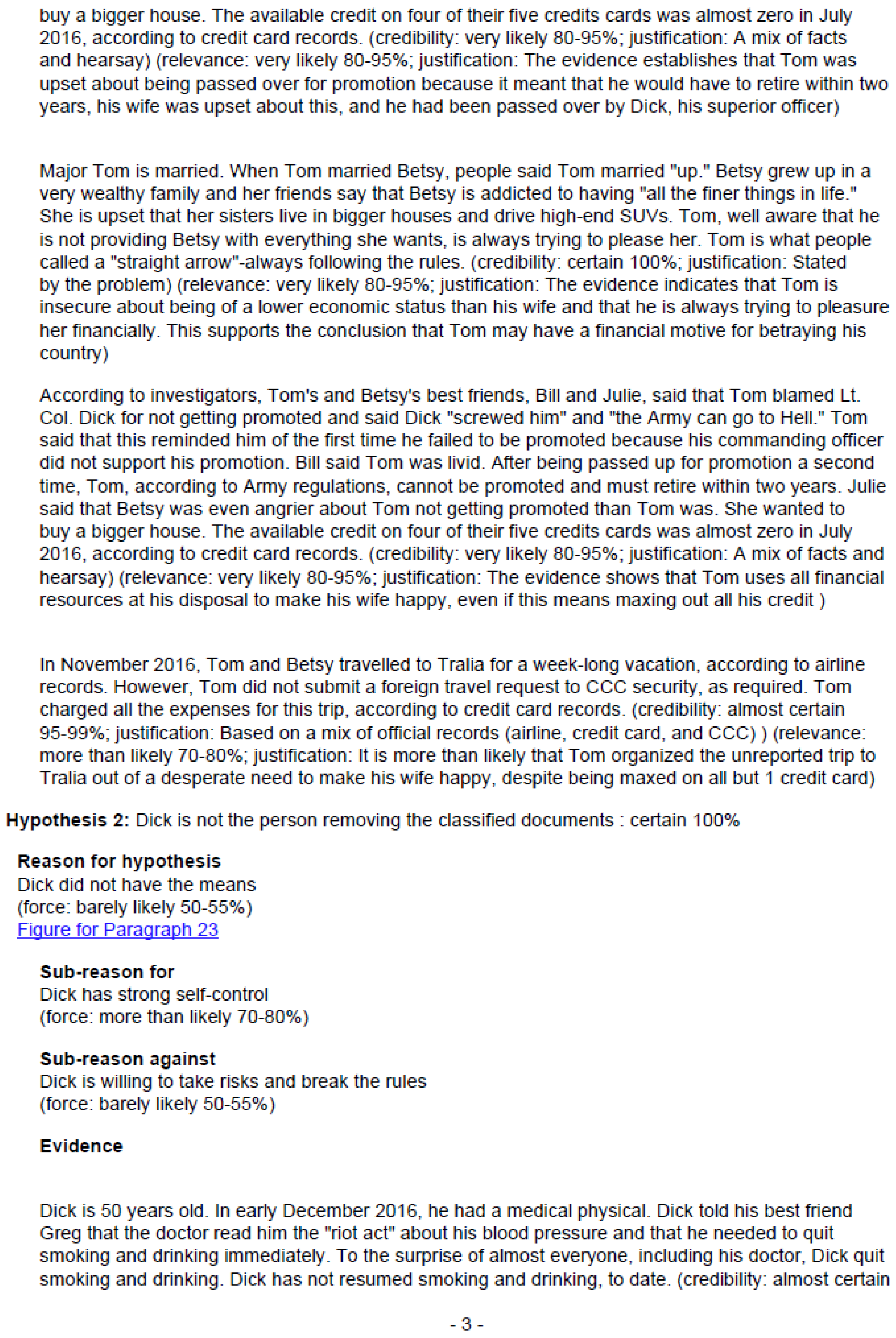
Sub-reason for
Tom was disgruntled at work
(force: likely 55-70%)

Sub-reason for
Tom was desperate for money to make his wife happy
(force: likely 55-70%)

Evidence

In September 2016, military records indicate that Tom was passed up for promotion for the second time. One of Dick's comments in Tom's personnel file was that Tom "did not live up to his potential and therefore should not be promoted." (credibility: certain 100%; justification: Military records) (relevance: very likely 80-95%; justification: The evidence shows that Tom was passed over for promotion potentially based on the recommendation of Dick)

According to investigators, Tom's and Betsy's best friends, Bill and Julie, said that Tom blamed Lt. Col. Dick for not getting promoted and said Dick "screwed him" and "the Army can go to Hell." Tom said that this reminded him of the first time he failed to be promoted because his commanding officer did not support his promotion. Bill said Tom was livid. After being passed up for promotion a second time, Tom, according to Army regulations, cannot be promoted and must retire within two years. Julie said that Betsy was even angrier about Tom not getting promoted than Tom was. She wanted to



buy a bigger house. The available credit on four of their five credits cards was almost zero in July 2016, according to credit card records. (credibility: very likely 80-95%; justification: A mix of facts and hearsay) (relevance: very likely 80-95%; justification: The evidence establishes that Tom was upset about being passed over for promotion because it meant that he would have to retire within two years, his wife was upset about this, and he had been passed over by Dick, his superior officer)

Major Tom is married. When Tom married Betsy, people said Tom married "up." Betsy grew up in a very wealthy family and her friends say that Betsy is addicted to having "all the finer things in life." She is upset that her sisters live in bigger houses and drive high-end SUVs. Tom, well aware that he is not providing Betsy with everything she wants, is always trying to please her. Tom is what people called a "straight arrow"-always following the rules. (credibility: certain 100%; justification: Stated by the problem) (relevance: very likely 80-95%; justification: The evidence indicates that Tom is insecure about being of a lower economic status than his wife and that he is always trying to pleasure her financially. This supports the conclusion that Tom may have a financial motive for betraying his country)

According to investigators, Tom's and Betsy's best friends, Bill and Julie, said that Tom blamed Lt. Col. Dick for not getting promoted and said Dick "screwed him" and "the Army can go to Hell." Tom said that this reminded him of the first time he failed to be promoted because his commanding officer did not support his promotion. Bill said Tom was livid. After being passed up for promotion a second time, Tom, according to Army regulations, cannot be promoted and must retire within two years. Julie said that Betsy was even angrier about Tom not getting promoted than Tom was. She wanted to buy a bigger house. The available credit on four of their five credits cards was almost zero in July 2016, according to credit card records. (credibility: very likely 80-95%; justification: A mix of facts and hearsay) (relevance: very likely 80-95%; justification: The evidence shows that Tom uses all financial resources at his disposal to make his wife happy, even if this means maxing out all his credit)

In November 2016, Tom and Betsy travelled to Tralia for a week-long vacation, according to airline records. However, Tom did not submit a foreign travel request to CCC security, as required. Tom charged all the expenses for this trip, according to credit card records. (credibility: almost certain 95-99%; justification: Based on a mix of official records (airline, credit card, and CCC)) (relevance: more than likely 70-80%; justification: It is more than likely that Tom organized the unreported trip to Tralia out of a desperate need to make his wife happy, despite being maxed on all but 1 credit card)

Hypothesis 2: Dick is not the person removing the classified documents : certain 100%

Reason for hypothesis
Dick did not have the means
(force: barely likely 50-55%)
Figure for Paragraph 23

Sub-reason for
Dick has strong self-control
(force: more than likely 70-80%)

Sub-reason against
Dick is willing to take risks and break the rules
(force: barely likely 50-55%)

Evidence

Dick is 50 years old. In early December 2016, he had a medical physical. Dick told his best friend Greg that the doctor read him the "riot act" about his blood pressure and that he needed to quit smoking and drinking immediately. To the surprise of almost everyone, including his doctor, Dick quit smoking and drinking. Dick has not resumed smoking and drinking, to date. (credibility: almost certain

95-99%; justification: Stated by the problem) (relevance: almost certain 95-99%; justification: Dick's determination to stay healthy by eliminating vices indicates a lack of impulsiveness)

Despite adultery being a criminal offense under Arboria's Uniform Code of Military Justice, Dick began an affair with Becky, a young woman, in early November 2016. According to investigators, Dick only told Greg Holland about the affair. Dick told Greg in late November that he likes Becky but he doesn't see the relationship going anywhere. He also told Greg that he "hated" the everyday pressure that Sally put on him regarding finances and that he was going to seek a divorce. Dick also said that Sally was one of the most superficial people he had ever met, and he doesn't know why he ever asked her to marry him. (credibility: almost certain 95-99%; justification: Stated by the problem) (relevance: almost certain 95-99%; justification: Dick was a cheater and was willing to break the law. However, this does not mean he had the capacity to commit treason.)

Dick loves to gamble. Every year Dick and his wife travel to Palisade, Arboria's gambling mecca in the mountains. Investigators discovered that in June 2016, Dick lost $30,000, and in October 2016, he hired a financial adviser who recommended that Dick take out a second-equity loan on his house, which Dick did. Relatives of Dick and Sally told investigators that Dick and Sally argue a lot about money. (credibility: almost certain 95-99%; justification: Stated by the problem) (relevance: almost certain 95-99%; justification: The evidence establishes that Dick has a higher than normal appetite for risk when it comes to gambling and is willing to leverage major personal assets against his gambling habit)

Reason for hypothesis
Dick did not have the motive
(force: barely likely 50-55%)
Figure for Paragraph 29

Sub-reason for
Dick would not have betrayed his country for his wife
(force: very likely 80-95%)

Sub-reason for
Dick did not feel aggrieved by being passed over for promotion
(force: very likely 80-95%)

Sub-reason against
Dick needed money
(force: barely likely 50-55%)

Evidence

Despite adultery being a criminal offense under Arboria's Uniform Code of Military Justice, Dick began an affair with Becky, a young woman, in early November 2016. According to investigators, Dick only told Greg Holland about the affair. Dick told Greg in late November that he likes Becky but he doesn't see the relationship going anywhere. He also told Greg that he "hated" the everyday pressure that Sally put on him regarding finances and that he was going to seek a divorce. Dick also said that Sally was one of the most superficial people he had ever met, and he doesn't know why he ever asked her to marry him. (credibility: certain 100%; justification: Stated by the problem) (relevance: very likely 80-95%; justification: The evidence shows that Dick was unhappy in his marriage and instead of succumbing to the financial pressure that his wife exerted on him, he was planning to divorce her.)

In October 2016, Dick was passed up for promotion to full colonel. The promotion would not have made Dick rich, but the extra money and perks would have eased some of the stress in making his loan and credit card obligations, allowing for more discretionary spending. Dick told Greg Holland, his best friend, that he was confident he would get promoted in the next promotion cycle. (credibility: very likely 80-95%; justification: Greg Holland was in a position to know Dick's true feelings about being passed over for promotion and has no reason to lie (Dick has solid alibis for two days when documents were passed to the Razmanian embassy and would not need help to make his defensive case) (relevance: almost certain 95-99%; justification: The evidence establishes that Dick was disgruntled despite being passed over for promotion)

Dick loves to gamble. Every year Dick and his wife travel to Palisade, Arboria's gambling mecca in the mountains. Investigators discovered that in June 2016, Dick lost $30,000, and in October 2016, he hired a financial adviser who recommended that Dick take out a second-equity loan on his house, which Dick did. Relatives of Dick and Sally told investigators that Dick and Sally argue a lot about money. (credibility: almost certain 95-99%; justification: A mix of facts and hearsay) (relevance: likely 55-70%; justification: It is likely that Dick needed money given his gambling habit and second equity loan)

Reason for hypothesis
Dick did not have the opportunity to deliver the documents to the Razmanian embassy (his access and ability to steal the documents in the first place is assumed based on the problem)
(force: certain 100%)
Figure for Paragraph 36

Sub-reason for
Dick has an alibi for 18 December
(force: certain 100%)

Sub-reason for
Dick has an alibi for 12 December
(force: certain 100%)

Evidence

According to CCC entry/exit records, Tom and Dick were in the office all day on 18 December. Harry called in sick. Paul, according to his company's records, reported to work at 11 AM on 18 December. No reason for his tardiness was specified. (credibility: almost certain 95-99%; justification: Based on official CCC records, though there is a small chance of recording error) (relevance: certain 100%; justification: The evidence establishes that Dick was at work for the full day on 18 December)

At around 10 AM on 18 December 2016, "Archimedes" tossed a message over the wall of our embassy in Monaca. This message included instructions for leaving $25,000 at a dead drop site along with additional instructions for receiving additional documents at another dead drop site.
The message again warned against communicating this information via electronic communication systems. (credibility: certain 100%; justification: Though the evidence is imprecise about the specific time, it is certain that the culprit threw over the documents around midday. This makes it impossible for any suspect who was in the office all day on any of the days in question to be the suspect.) (relevance: certain 100%; justification: The evidence shows that the drop occurred during the day on 18 December)

According to CCC entry/exit records, Tom and Dick were in the office all day on 12 December. Harry was out of the building from 10 AM to 1 PM that day. Paul called in sick on 12 December, according to his company's records. (credibility: almost certain 95-99%; justification: Based on official CCC records, though there is a small chance of recording error) (relevance: certain 100%; justification: The evidence establishes that Dick was at work for the full day on 12 December)

At around noon on 12 December 2016, the same individual tossed a copy of an Arborian intelligence document on Plumistan's new tank over the wall of our embassy in Monaca. Archimedes asked that this information not be communicated back to Razmania via electronic communications but be hand carried back (in a diplomatic pouch that cannot be opened or otherwise searched) to Razmania. Taped to this document was a message that again noted Archimedes' concern about Razmania's security but also offered long-term cooperation with Razmania in exchange for money. (credibility: certain 100%; justification: Stated by the problem) (relevance: certain 100%; justification: The evidence shows that the drop occurred during the day on 12 December)

Hypothesis 3: Harry is the person removing the classified documents : almost certain 95-99%

Reason for hypothesis
Harry had the means
(force: likely 55-70%)
Figure for Paragraph 44

Sub-reason for
Harry has had suspicious contact with suspected Razmanian intelligence before and after the document theft
(force: likely 55-70%)

Sub-reason for
Harry has no close personal connections in Arboria
(force: barely likely 50-55%)

Sub-reason for
Harry had special access to valuable (sensitive and timely) information about Plumistan
(force: more than likely 70-80%)

Evidence

On 10 July 2016, Harry had dinner at the Great Inn in downtown Monaca with Dr. Mark Down, who is listed as a Cultural Attache in the Razmanian Embassy. According to CCC security records, Harry reported this contact. Arboria's internal security agency suspects that Down's real employer is the MIS and that his cultural-attache title is just a cover. However, this is not known with certainty. In his report, Harry noted that he met Down during one of his visits to Razmania, and they both share a deep interest in Razmania's history. (credibility: certain 100%; justification: Harry reported the meeting himself and Arborian intelligence logged the report) (relevance: very likely 80-95%; justification: It is not certain that Mark Downs is actually Razmanian intelligence but it is very likely given the suspicion)

Arboria's internal security agency reviewed the credit card records of Harry and Dr. Mark Down in February 2017, and discovered that both Harry and Down paid bills at the Great Inn on 15 October 2016 within several minutes of each other. Harry did not report any contact with Down on or near that date, according to CCC security records. At this time, it is uncertain whether such contact occurred. (credibility: likely 55-70%; justification: The credibility of this evidence is likely given the prior meeting with Harry) (relevance: likely 55-70%; justification: A second suspicious instance of possible contact between Harry and a suspected Razmanian intelligence officer after the document theft hints at a quid pro quo arrangement)

Captain Harry is single and has never been married. He lives alone and has no close friends. People who have worked with Harry believe he is a bit of a recluse. Some of this is due to Harry having Asperger's Syndrome, and some of this is due to Harry's experience growing up. Harry's parents were killed in a car accident when Harry was 18 months old, and he moved from foster home to foster home during the first ten years of his life. A 2010 psychological assessment of Harry (conducted when Harry was commissioned as an army officer) noted that Harry has "trust"" issues. (credibility: certain 100%; justification: Stated in the problem) (relevance: very likely 80-95%; justification: Harry's life history, past psychological assessment and current behaviors almost certain establish that he has no close personal connections in Arboria)

In late September 2016, Plumistan began to use a new encryption system for its diplomatic messages. Harry was temporarily assigned to the team that was tasked with breaking this new encryption system. The system was broken in early November after team members-including Harry-worked several 65-hour-work weeks running computer program after computer program. (credibility: certain 100%; justification: Stated in the problem) (relevance: almost certain 95-99%; justification: The work on breaking Plumistan's encryption protocols almost certainly gave Harry access to current and valuable information on Plumistan)

Reason for hypothesis
Harry had the motive
(force: likely 55-70%)
Figure for Paragraph 52

Sub-reason for
Harry has a deep emotional attachment to Razmania and cared about issues that impacted its safety and security : certain 100%
Plumistan is an existential threat to Razmania : certain 100%
Razmania wants more intelligence from Arboria : certain 100%
(force: very likely 80-95%)

Evidence

Harry's mother was born and raised in Razmania. In office conversations, Harry is quite vocal about Razmania's geopolitical vulnerability and is adamant that Arboria is not doing enough to defend Razmania. Most people in Harry's office no longer talk about Razmania in Harry's presence because of Harry's emotional outbursts during these discussions. Harry has traveled to Razmania at least once a year for the past seven years. He has reported all of these trips to the Office of Security in C/M-2. (credibility: certain 100%; justification: Given by problem) (relevance: certain 100%; justification: The evidence establishes with certainty that Harry has deep and abiding attachments to the country of his mother's birth)

On 10 July 2016, Harry had dinner at the Great Inn in downtown Monaca with Dr. Mark Down, who is listed as a Cultural Attache in the Razmanian Embassy. According to CCC security records, Harry reported this contact. Arboria's internal security agency suspects that Down's real employer is the MIS and that his cultural-attache title is just a cover. However, this is not known with certainty. In his report, Harry noted that he met Down during one of his visits to Razmania, and they both share a deep interest in Razmania's history. (credibility: almost certain 95-99%; justification: The evidence is based on Harry's own reports to Arborian intelligence) (relevance: almost certain 95-99%; justification: The evidence establishes Harry's interest in Razmanian history)

Razmania's main threat is Plumistan, a large country that borders it to the north. Plumistan has publicly acknowledged that its main strategic goal is the annihilation of Razmania. Plumistan has also threatened to support terrorist attacks against Arboria. (credibility: certain 100%; justification: Given by problem) (relevance: certain 100%; justification: Stated in evidence E4)

Arboria and Razmania are allies of sorts, but not all of their interests coincide. As a result, Arboria does not share all of its intelligence with Razmania. Thus, Razmania's Intelligence Service (MIS) actively tries to recruit employees of Arboria's CCC as spies. Unbeknownst to Razmania, Arboria is able to break the diplomatic encryption codes used in Razmania's electronic communications. (credibility: certain 100%; justification: Given by problem) (relevance: certain 100%; justification: Stated in evidence E3)

Reason for hypothesis

Harry had the opportunity to deliver the documents to the Razmanian embassy (his ability to access and steal the documents in the first place is assumed based on the problem)
(force: almost certain 95-99%)
Figure for Paragraph 58

Sub-reason for

Harry has no alibi for 12 December
(force: very likely 80-95%)

Sub-reason for

Harry has no alibi for 18 December
(force: very likely 80-95%)

Sub-reason for

On 9 December Harry dropped off the documents at the Razmanian embassy "around noon" and still made it to work by 12 p.m.
(force: almost certain 95-99%)

Evidence

According to CCC entry/exit records, Tom and Dick were in the office all day on 12 December. Harry was out of the building from 10 AM to 1 PM that day. Paul called in sick on 12 December, according to his company's records. (credibility: almost certain 95-99%; justification: Official CCC records. A miniscule chance of error) (relevance: very likely 80-95%; justification: The evidence establishes that Harry was not at work at the time of the drop of 12 December.)

At around noon on 12 December 2016, the same individual tossed a copy of an Arborian intelligence document on Plumistan's new tank over the wall of our embassy in Monaca. Archimedes asked that this information not be communicated back to Razmania via electronic communications but be hand carried back (in a diplomatic pouch that cannot be opened or otherwise searched) to Razmania. Taped to this document was a message that again noted Archimedes' concern about Razmania's security but also offered long-term cooperation with Razmania in exchange for money. (credibility: certain 100%; justification: Stated in the problem) (relevance: very likely 80-95%; justification: The evidence establishes the time of the drop at the Razmanian embassy during the day of 12 December)

According to CCC entry/exit records, Tom and Dick were in the office all day on 18 December. Harry called in sick. Paul, according to his company's records, reported to work at 11 AM on 18 December. No reason for his tardiness was specified. (credibility: almost certain 95-99%; justification: Official CCC records. A miniscule chance of error) (relevance: very likely 80-95%; justification: The evidence establishes that Harry was not at work during the rough window of the drop on 18 December)

At around 10 AM on 18 December 2016, "Archimedes" tossed a message over the wall of our embassy in Monaca. This message included instructions for leaving $25,000 at a dead drop site along with additional instructions for receiving additional documents at another dead drop site.
The message again warned against communicating this information via electronic communication systems. (credibility: certain 100%; justification: Stated in the problem) (relevance: very likely 80-95%; justification: The evidence establishes that a drop occurred during the day on 18 December)

At around noon on 9 December 2016, an individual identifying themselves as "Archimedes" tossed a copy of an Arborian intelligence document over the wall of our embassy in Monaca. The document was entitled "Military Encryption Codes of Plumistan" and the document was identified as No. 87 out of 200. Archimedes said they were a friend of Razmania, and the document enclosed was meant to show what kinds of information they could provide. (credibility: very likely 80-95%; justification: Stated in the problem. However, the imprecision of the time frame given ("around noon") weakens the credibility of this piece of evidence specifically in relation to this hypothesis) (relevance: almost certain 95-99%; justification: The time frame given in the evidence is not precise enough to eliminate the possibility of someone making the drop "around noon" and getting to work by 12 p.m.)

According to CCC entry/exit records, Tom and Harry were in the office at noon on 9 December 2016. Entry/exit records show that Dick was out of the office from 11AM to 1PM on this day. Paul Jones was in CCC's headquarters from 8AM to 11AM on 9 December, according to entry and exit logs for CCC's headquarters. (credibility: almost certain 95-99%; justification: The evidence establishes that Harry arrive at work at 12 p.m.) (relevance: almost certain 95-99%)

Hypothesis 4: Paul is not the person removing the classified documents : very likely 80-95%

Reason for hypothesis
Paul did not have the means
(force: likely 55-70%)
Figure for Paragraph 69

Sub-reason for
Paul has lived a clean life for decades
(force: likely 55-70%)

Sub-reason for
Paul lacks the geopolitical awareness to know the contextual value of the documents that were stolen
(force: likely 55-70%)

Evidence

Paul, age 45, has had a rough life. His father was abusive and his mother was a cocaine addict. He dropped out of high school in the 10th grade. As a young man in his early twenties, Paul committed a number of felonies, including armed robbery, according to state penal records. Paul spent four years in prison. While in prison, Paul found God and turned his life around. He has not had a speeding ticket in 20 years. When not working, Paul spends most of his day watching reality TV. A friend of Paul's once said that Paul probably could not name more than three capitals in the world, not counting Monaca. (credibility: certain 100%; justification: Stated in the problem) (relevance: almost certain 95-99%; justification: The evidence establishes the fact that Paul has lived a reformed life, up to and including strict adherence to traffic laws, in 20 years)

Paul, age 45, has had a rough life. His father was abusive and his mother was a cocaine addict. He dropped out of high school in the 10th grade. As a young man in his early twenties, Paul committed a number of felonies, including armed robbery, according to state penal records. Paul spent four years in prison. While in prison, Paul found God and turned his life around. He has not had a speeding ticket in 20 years. When not working, Paul spends most of his day watching reality TV. A friend of Paul's once said that Paul probably could not name more than three capitals in the world, not counting Monaca. (credibility: certain 100%; justification: Stated in the problem) (relevance: more than likely 70-80%; justification: Paul's personal history and current behaviors suggest he does not have significant geopolitical awareness)

Reason for hypothesis
Paul did not have the opportunity to take the documents

(force: almost certain 95-99%)
Figure for Paragraph 74

Sub-reason for
Direct evidence:
An uncleared custodian named Paul Jones enters 3C12 every workday to empty the trash cans that are only authorized for unclassified material and rubbish. Because there are mounds of classified material everywhere, Harry, Tom, or Dick always escort Paul when he is in the vault and monitor his movements closely. They are careful not to discuss classified material when Paul is in the vault. At the end of the day, the vault is physically locked and an electronic security system is activated to ensure that there is no unregistered entry. No one other than Paul has partial access to 3C12. (credibility: certain 100%; justification: Stated in the problem) (relevance: almost certain 95-99%; justification: Paul could not have taken the documents without the cooperation of whichever employee (Tom, Dick or Harry) escorted him on his rubbish collection rounds. Because we are told to assume there was no cooperation, Paul could not have taken the documents)
(force: almost certain 95-99%)

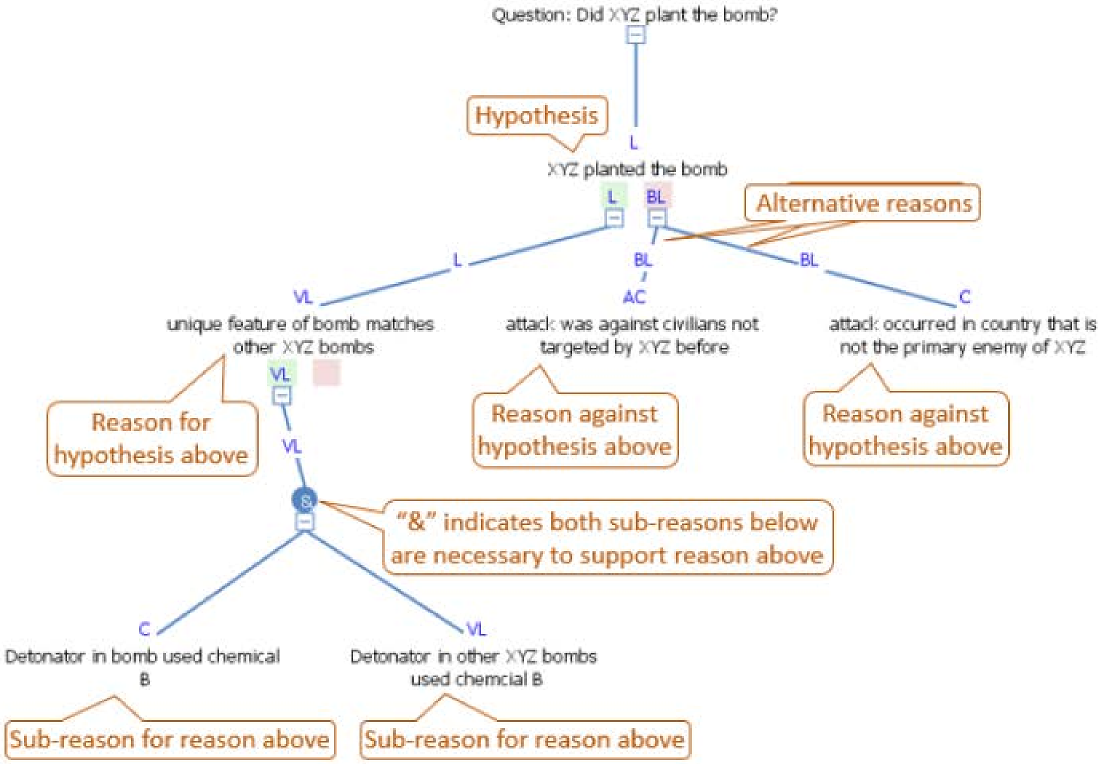
Reason against hypothesis
Paul had the motive
(force: likely 55-70%)
Figure for Paragraph 76

Sub-reason for
Paul was desperate for funds to pay for his daughter's liver transplant
(force: almost certain 95-99%)

Evidence

Paul has an older daughter, named Jennifer, who he adores. Jennifer needs a liver transplant, but does not have insurance. The cost for the transplant is about $100,000. Paul is not sleeping well because of his daughter's condition. (credibility: certain 100%) (relevance: almost certain 95-99%; justification: A liver transplant is a serious and expensive medical necessity, and Paul is understandably stressed about his daughter needing one)

Review the Explanation of this Sample Analysis to Understand the Reasoning Diagrams in the Appendix


Question: Did XYZ plant the bomb?
L
XYZ planted the bomb
L
BL
L
VL
unique feature of bomb matches other XYZ bombs
BL
AC
attack was against civilians not targeted by XYZ before
BL
C
attack occurred in country that is not the primary enemy of XYZ
VL
VL
&
C
Detonator in bomb used chemical B
VL
Detonator in other XYZ bombs used chemcial B

Hypothesis

Alternative reasons

Reason for hypothesis above

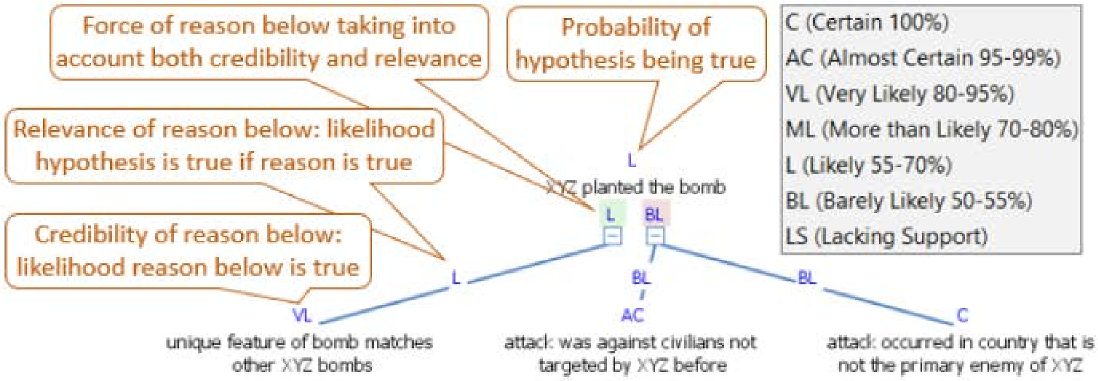
Reason against hypothesis above

Reason against hypothesis above

"&" indicates both sub-reasons below are necessary to support reason above

Sub-reason for reason above

Sub-reason for reason above


L
XYZ planted the bomb
L
BL
L
VL
unique feature of bomb matches other XYZ bombs
BL
AC
attack was against civilians not targeted by XYZ before
BL
C
attack occurred in country that is not the primary enemy of XYZ

Force of reason below taking into account both credibility and relevance

Probability of hypothesis being true

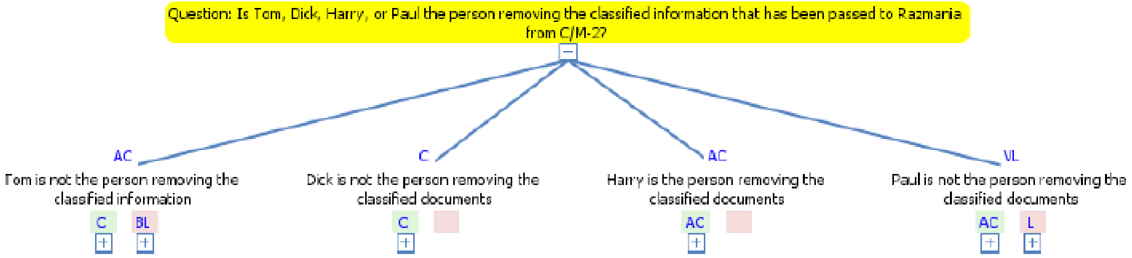
Relevance of reason below: likelihood hypothesis is true if reason is true

Credibility of reason below: likelihood reason below is true

C (Certain 100%)
AC (Almost Certain 95-99%)
VL (Very Likely 80-95%)
ML (More than Likely 70-80%)
L (Likely 55-70%)
BL (Barely Likely 50-55%)
LS (Lacking Support)

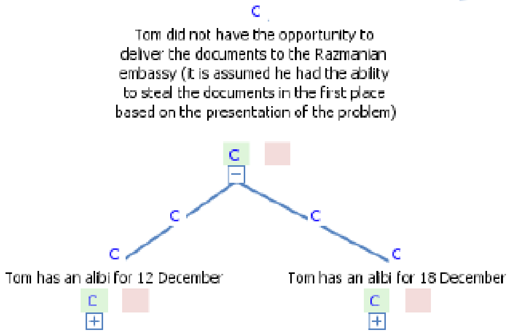
Appendix: Is Tom, Dick, Harry, or Paul the person removing the classified information that has been passed to Razmania from C/M-2?

Figure for Paragraph 1:


Question: Is Tom, Dick, Harry, or Paul the person removing the classified information that has been passed to Razmania from C/M-2?
AC
Tom is not the person removing the classified information
C
BL
C
Dick is not the person removing the classified documents
C
AC
Harry is the person removing the classified documents
AC
VL
Paul is not the person removing the classified documents
AC
L

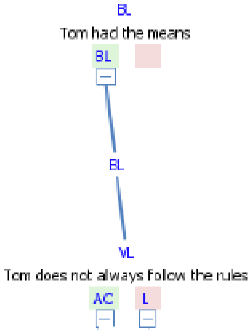
Return to Report

Figure for Paragraph 3:


C
Tom did not have the opportunity to deliver the documents to the Razmanian embassy (it is assumed he had the ability to steal the documents in the first place based on the presentation of the problem)
C
C
C
C
C
Tom has an alibi for 12 December
C
Tom has an alibi for 18 December
C

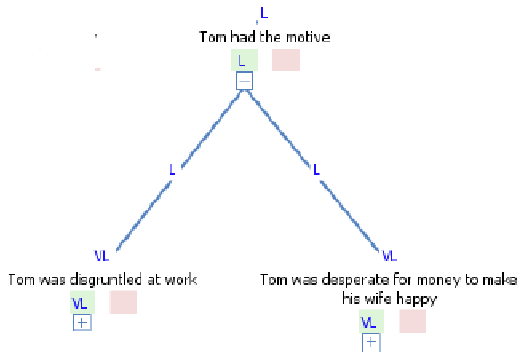
Return to Report

Figure for Paragraph 10:



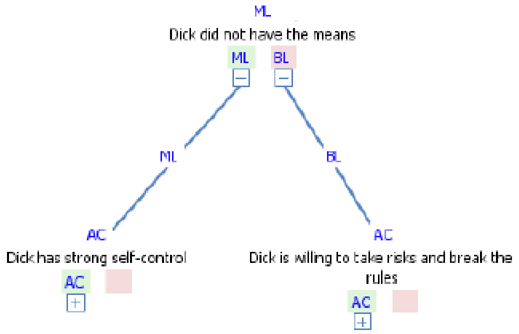
Return to Report

Figure for Paragraph 14:


L
Tom had the motive
L
L
L
VL
Tom was disgruntled at work
VL
VL
Tom was desperate for money to make his wife happy
VL

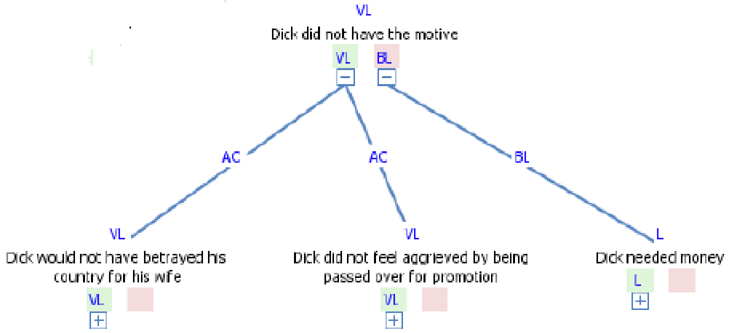
Return to Report

Figure for Paragraph 23:


ML
Dick did not have the means
ML
BL
ML
BL
AC
Dick has strong self-control
AC
AC
Dick is willing to take risks and break the rules
AC

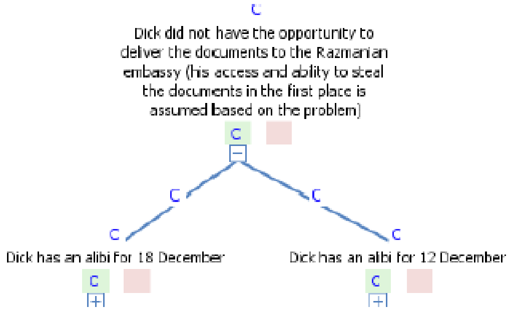
Return to Report

Figure for Paragraph 29:


VL
Dick did not have the motive
VL
BL
AC
AC
BL
VL
Dick would not have betrayed his country for his wife
VL
VL
Dick did not feel aggrieved by being passed over for promotion
VL
L
Dick needed money
L

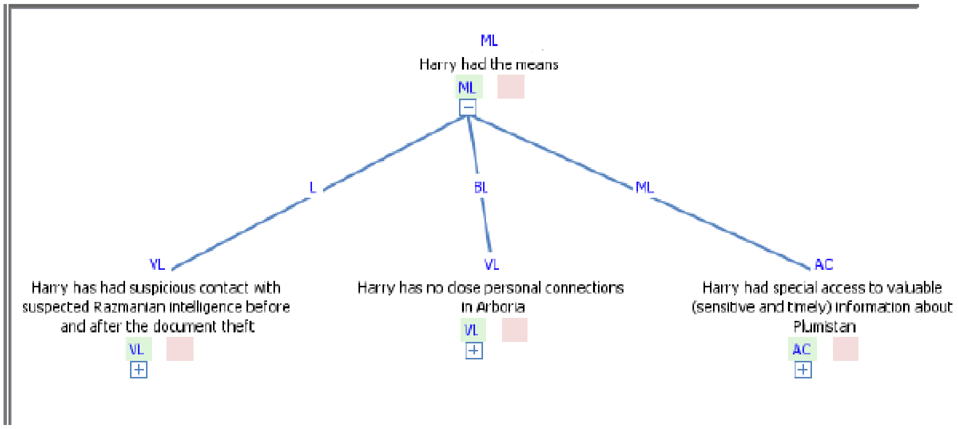
Return to Report

Figure for Paragraph 36:


C
Dick did not have the opportunity to deliver the documents to the Razmanian embassy (his access and ability to steal the documents in the first place is assumed based on the problem)
C
C
C
C
Dick has an alibi for 18 December
C
C
Dick has an albi for 12 December
C

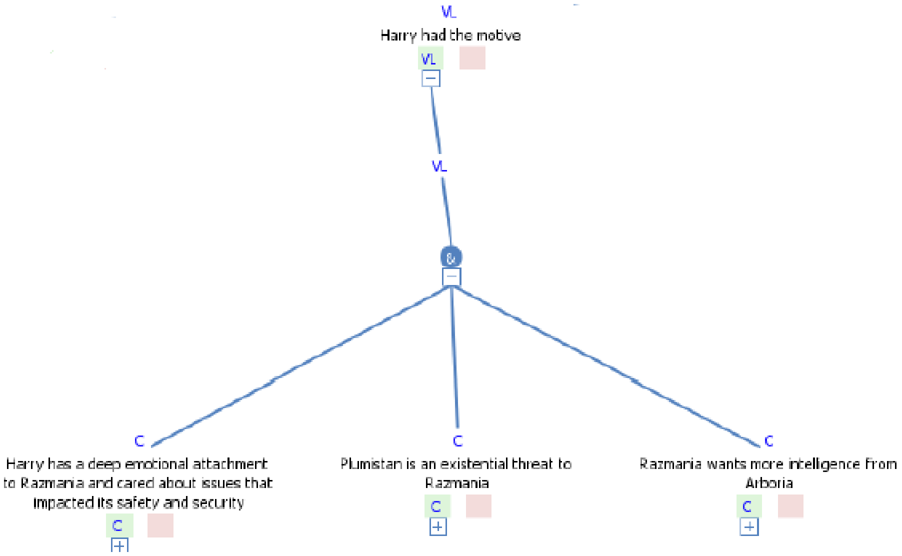
Return to Report

Figure for Paragraph 44:


ML
Harry had the means
ML
L
BL
ML
VL
Harry has had suspicious contact with suspected Razmanian intelligence before and after the document theft
VL
VL
Harry has no close personal connections in Arboria
VL
AC
Harry had special access to valuable (sensitive and timely) information about Plumistan
AC

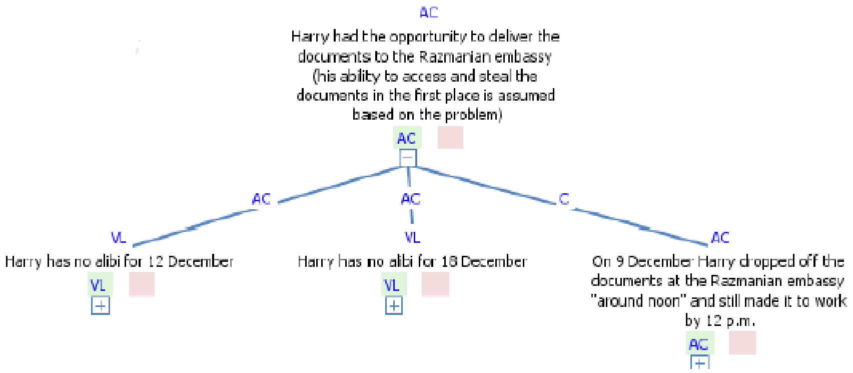
Return to Report

Figure for Paragraph 52:


VL
Harry had the motive
VL
VL
&
C
Harry has a deep emotional attachment to Razmania and cared about issues that impacted its safety and security
C
C
Plumistan is an existential threat to Razmania
C
C
Razmania wants more intelligence from Arboria
C

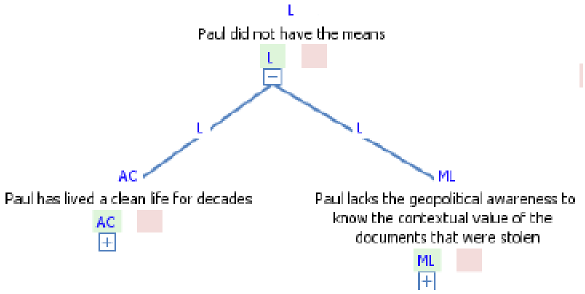
Return to Report

Figure for Paragraph 58:


AC
Harry had the opportunity to deliver the documents to the Razmanian embassy (his ability to access and steal the documents in the first place is assumed based on the problem)
AC
AC
AC
C
VL
Harry has no alibi for 12 December
VL
VL
Harry has no alibi for 18 December
VL
AC
On 9 December Harry dropped off the documents at the Razmanian embassy "around noon" and still made it to work by 12 p.m.
AC

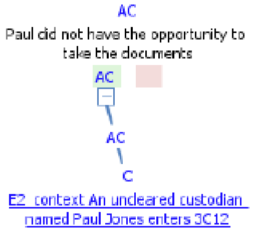
Return to Report

Figure for Paragraph 69:


L
Paul did not have the means
L
L
L
AC
Paul has lived a clean life for decades
AC
ML
Paul lacks the geopolitical awareness to know the contextual value of the documents that were stolen
ML

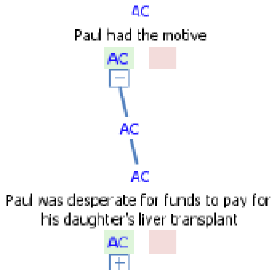
Return to Report

Figure for Paragraph 74:


AC
Paul did not have the opportunity to take the documents
AC
AC
C
E2 context An uncleared custodian named Paul Jones enters 3C12

Return to Report

Figure for Paragraph 76:


AC
Paul had the motive
AC
AC
AC
Paul was desperate for funds to pay for his daughter's liver transplant
AC

Return to Report