Computer Supported Cooperative Work (2008) 17:353–393
DOI 10.1007/s10606-008-9080-9

The CACHE Study: Group Effects in Computer-supported Collaborative Analysis

Gregorio Convertino[1], Dorrit Billman[2,5], Peter Pirolli[2], J. P. Massar[3] & Jeff Shrager[4]
[1]*College of Information Sciences and Technology, Penn State University, University Park, PA 16802, USA (E-mail: gconvertino@gmail.com);* [2]*Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA (E-mail: pirolli@parc.com);* [3]*Berkeley, CA, USA;* [4]*Symbolic Systems Program, Stanford University, Stanford, CA 94305, USA (E-mail: jshrager@stanford.edu);* [5]*Center for Study of Language and Information, Stanford University, Cordura Hall, Rm 110, Stanford, CA 94305, USA (E-mail: billman@psych.stanford.edu)*

Abstract. The present experiment investigates effects of group composition in computer-supported collaborative intelligence analysis. Human cognition, though highly adaptive, is also quite limited, leading to systematic errors and limitations in performance – that is, *biases*. We experimentally investigated the impact of group composition on an individual's bias, by composing groups that differ in whether their members initial beliefs are diverse (heterogeneous group) or similar (homogeneous group). We study three-member, distributed, computer-supported teams in heterogeneous, homogeneous, and solo (or nominal) groups. We measured bias in final judgment, and also in the selection and evaluation of the evidence that contributed to the final beliefs. The distributed teams collaborated via CACHE-A, a web-based software environment that supports a collaborative version of Analysis of Competing Hypotheses (or ACH, a method used by intelligence analysts). Individuals in Heterogeneous Groups showed no net process cost, relative to noninteracting individuals. Both heterogeneous and solo (noninteracting) groups debiased strongly, given a stream of balanced evidence. In contrast, individuals in Homogenous Groups did worst, accentuating their initial bias rather than debiasing. We offer suggestions about how CACHE-A supports collaborative analysis, and how experimental investigation in this research area can contribute to design of CSCW systems.

Key words: CACHE, collaboration, intelligence analysis, CSCW, group bias, group decision-making

1. Introduction

The need for better ways to support intelligence analysis is clear from recent events (e.g., The United States 9/11 Commission 2004). Intelligence analysis is made particularly difficult both by the intrinsic complexity of the task and by the cognitive limitations of analysts. Especially problematic are biases inherent in human reasoning. Both collaboration and technology are often suggested as ways

of enhancing human judgment in complex decisions (e.g., Johnston 2003 on teamwork; Card 2005 on external aids). However, the structures of collaboration can vary significantly, and it is not known which collaborative structures are most effective for debiasing, particularly in the context of computer-supported intelligence analysis.

Our work investigates the impact of group structure on debiasing judgment and is motivated by studies of individuals and groups over several decades (e.g., Kahneman et al. 1982; Heuer 1999; Kerr and Tindale 2004). These studies have provided a richer understanding of bias (systematic departures from an ideal standard) that occurs in human judgment and decision-making under uncertainty. This understanding informs research in computer-supported cooperative work (CSCW): design of CSCW systems benefits from an understanding of the component processes and their outcomes, at an individual and group level. The present experiment exploits a novel CSCW system that we developed. This system is used as a test bed for studying the effects of group composition on the quality of the intelligence analysis. We also evaluate the use of this prototyped collaborative software, which was designed to help analysts leverage one another's reasoning and fluently support transitions among varying work configurations.

In the next sections, we review literature on bias in decision-making and we identify key factors that make collaborative analysis a complex task. We then present an experiment that investigates one aspect of collaborative analysis: *How group composition affects the performance of group members*. Participants work in a computer environment that is a simplified version of a collaborative analysis system called CACHE (Collaborative Analysis of Competing Hypotheses Environment; Shrager 2005). We investigate the impact of group composition on bias and on coverage of relevant information. We also comment on system usability and design.

2. The study of bias

In this section we organize the research on bias by whether the individual or group is the focus and by whether or not the focus is on supporting tools, as shown schematically in Figure 1. We first consider research on biased cognition by unaided individuals, and then discuss the efforts that have been directed toward enhancing the cognition of individuals through methods or technology. Then we address whether and when a group improves performance over individuals on challenging cognitive tasks. Finally, we consider recent research that attempts to enhance performance of groups through methods or technology.

2.1. Bias in individuals

Human cognition, though highly adaptive, is also quite limited, leading to systematic errors and limitations in performance – that is, *biases*. Psychologists have demonstrated response patterns indicating several forms of bias, including

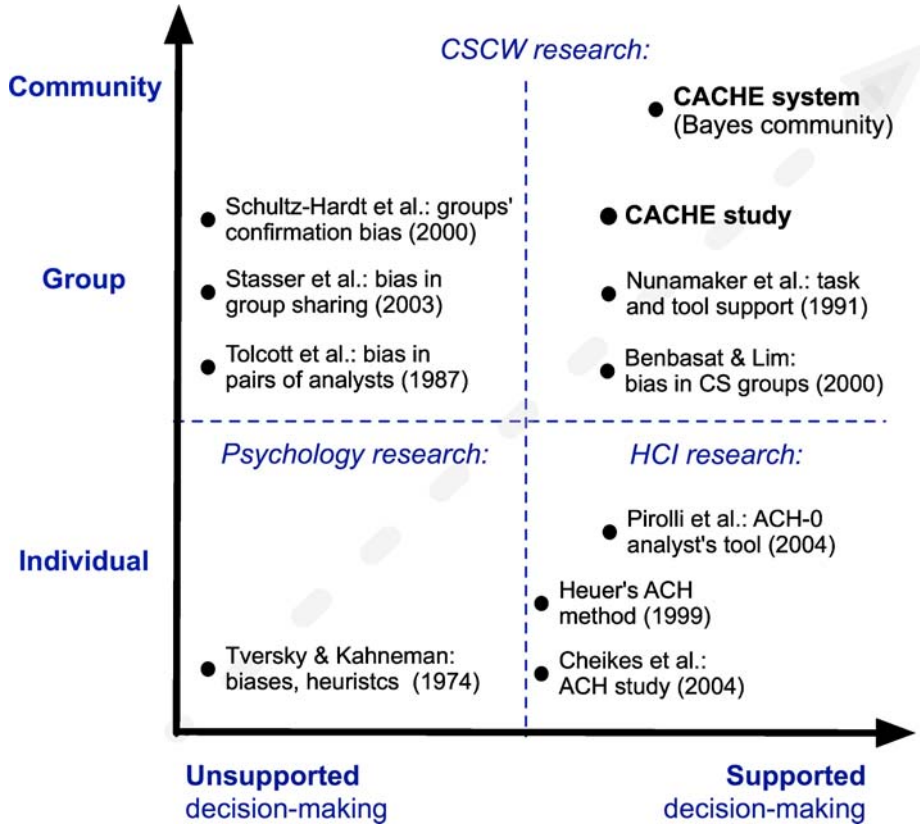

Figure 1 Map of the research space on bias by level of analysis and support. The *Y*-axis indicates the size of the decision-making unit, from individual to large community. The *X*-axis indicates the degree of support for decision-making, from unsupported to supported by both complex systems and methods. Prior literature most relevant to our work ("CACHE study," *top right*) is mapped onto this research space.

representativeness, availability, framing, and confirmation biases (e.g., Kahneman et al. 1982; Wason 1960). Furthermore, errors resulting from biases that occur in a single judgment or decision may be amplified by chained inferences, where the output of one decision serves as input to the next (Gettys et al. 1982). For example, Johnson and Halpin (1974) working in an intelligence domain found that estimates became increasingly extreme, relative to a Bayesian standard, as the number of intermediate inferences increased.

An important source of bias pertains to time and order of information. People tend to over-commit to explanations, hypotheses, or choices that are encountered or preferred early in the process. Various factors can contribute to overly strong commitment to prior beliefs. Researchers have shown that people tend to hold too strongly to an initial value, which acts as a reference point (an anchor); then the subsequent adjustments made with respect to this point are typically insufficient

to fully account for new information (anchoring heuristic, Tversky and Kahneman 1974). People may also expend too little effort in generating alternative possibilities (Heuer 1999). The selection or processing of new evidence to assess a belief may favor an initial, preferred option (Wason 1960; Klayman and Ha 1987). Finally, the integration of information selected to reach a decision may be biased (Benjamin 1990). In this paper, we will refer broadly to overly strong commitment to an early favorite as *confirmation bias*.

Understanding biases is particularly important if the biases result in significant departures from sound (*normative*) reasoning. A key issue is the identification of normative outcomes or procedures to which human judgment can be compared. Identification of the normative outcome or procedure is often difficult. For example, how strongly should a scientist stick with a current theory, explaining away surprising results as due to methodological problems (Fugelsang et al. 2004)? How strongly should an analyst hold to accepted geopolitical explanations (Johnston 2005)? As a consequence, most research has been done in the laboratory with simple materials and tasks which can be more easily controlled. One of the methods for assessing bias is whether judgments are influenced by factors that are normatively irrelevant to the task. For example, in many contexts the order in which an individual personally encounters evidence is irrelevant to its importance; critical facts may turn up early or late in a case. Thus, if a piece of evidence is encountered earlier, it should *not* be weighted more or less than if encountered later. Order of presenting evidence in fact produces a variety of order effects (Chapman et al. 1996; Kerstholt and Jackson 1998; for a review see Hogarth and Einhorn 1992). For example, heavy reliance on early evidence is likely for long sequences of evidence.

Researchers studying naturalistic decision-making have questioned both the relevance of traditional characterizations of bias and whether cognitive biases identified in laboratory tasks negatively affect the performance of actual experts or skilled workers in their own domain of expertise (Lipshitz et al. 2001). Indeed, in the domain of military intelligence Tolcott et al. (1989) suggest that biases may be lessened for more expert decision makers. While the extent and nature of bias by experts certainly varies by domain (Shanteau 1992) and task, expert performance remains biased in many circumstances. Of relevance is, specifically, the biased performance identified in a variety of professional tasks that require reflection on and *analysis of extensive information in circumstances with considerable novelty and change* (e.g. for auditors Davis and Ashton 2002; for medical doctors Bornstein and Emler 2001). In addition, researchers have sought to create laboratory experiments that are analogs of work tasks by placing domain practitioners in settings with realistic materials and activities. For example, researchers (Adelman et al. 1993; Perrin et al. 2001) have found effects of presentation order on judgments of air defense personnel working on a naturalistic plane classification task. These studies provide complementary

support for concerns raised in professional practice based on broader but less formal evidence (e.g., Heuer 1999; Johnston 2003, 2005).

Experts have extensive knowledge and entrenched beliefs. This is usually helpful, but may itself create bias through over-commitment to these beliefs (the paradox of expertise, Camerer and Johnson, 1991; Johnston 2005). Further, experts in many domains are faced with novel tasks, much potentially relevant information, high expectations for performance, and ambiguous performance criteria. This is often the case for intelligence analysts. These work conditions may encourage shortcuts and restrict careful analysis. Indeed, practitioners may lack procedures that are both effective and unbiased.

In sum, confirmation bias in our broad sense is widespread; it occurs in a wide range of settings including those with high stakes, and both laypeople and professionals are vulnerable. Thus, developing methods and tools that support effective and unbiased performance is an important goal for CSCW research and design.

2.2. Bias in individuals supported by methods or technology

Technology support offers one approach to improving professional judgment in the face of the many challenges to unbiased thinking. This may be as simple as a checklist to support memory in critical aviation tasks. However, support is often more complex than that, and many computerized decision support systems have been developed to aid various task components (for a review Keefer et al. 2004; particularly on bias Arnott 2006). Many of these tools rely on the user's ability to formalize and quantify judgments, and often focus on a narrow aspect of a reasoning task, such as reminding or explicit value elicitation. Moreover, although there are many decision support systems available (see George et al. 2000 on bias; Keefer et al. 2004 on decision analysis), virtually none supports the full set of functions needed by analysts. In this respect, Scholtz et al. (2006) discuss some of the methodological challenges when collections of software tools are developed and assessed, for intelligence analysis. However, even with technological support the performance by individuals performing challenging cognitive tasks is still limited and biased, and these problems are particularly acute in intelligence analysis.

Responding to problems in intelligence analysis, Heuer (1999) not only identified cognitive biases that impair performance, but also made a wide set of recommendations. He particularly defines a semi-formal method that he calls ACH – Analysis of Competing Hypotheses. ACH is based upon the classical method of decision matrices. The analyst develops a full set of alternative (competing) hypotheses, which form the columns of a decision matrix, and the available evidence which forms the rows (Figure 2, Section 5.2). Each cell of the resulting matrix represents the relation between a piece of evidence, and a particular hypothesis.

Figure 2 Alpha's ACH matrix. Three hypotheses are listed in the column headings. The pieces of evidence are listed along the rows. The cells indicate consistent and inconsistent values for the relations between a piece of evidence and a hypothesis.

To reduce bias, analysts trained in the ACH methodology are supposed to focus on *inconsistent evidence* rather than on consistent evidence, and on *diagnostic evidence* – which maximally discriminates hypotheses from one another – rather than on uninformative evidence. Moreover, Heuer encourages analysts to evaluate the *relation of a piece of evidence* (row) *to all the alternative hypotheses* (columns). The ACH method was investigated by Cheikes et al. (2004) who found that intelligence analysts working alone showed less bias favoring their initial expectation when they used the ACH method than when using their default methods.

ACH0 (Pirolli et al. 2005) is a desktop reasoning tool that supports a significant portion of the ACH method. It provides an ACH Hypotheses-by-Evidence matrix with cells that indicate the relationships between evidence and hypotheses. ACH0 provides automatic computer integration of the user-entered cell values to derive a support measure for each alternative hypothesis. This tool has been adopted and used by a part of the intelligence community.

ACH0 may benefit analysts by providing an easy-to-use, external representation that can be shared by individuals working on the same or related problems. Pirolli et al. (2005) studied the ACH method used by individuals and compared performance using a prototype of the ACH0 software versus paper and pencil ACH analysis. Pirolli, et al. asked US Navy post-graduate student analysts to analyze a simulated intelligence problem, and found trends, but not significant differences between the two platforms (software ACH vs. paper and pencil ACH). The CACHE-A system, used in the experiment reported in this paper, extends the

ACH0 individual prototype to support collaborative intelligence analysis (see description of CACHE-A in Section 5).

2.3. Bias in groups

One route to improving performance on challenging cognitive tasks is to form groups of individuals working together. Indeed, many complex cognitive tasks can only be solved when people work together. However, groups introduce their own coordination and process costs (i.e., *group process losses*) which cut into a group's potential performance (e.g., on bias, Kerr et al. 1996; on process losses review by Kerr and Tindale 2004). For example, one member cannot talk while also listening to another, leading to production blocking (e.g., Gallupe et al. 1991).

Group work not only incurs group costs, but also bias. One source of bias in groups is persistence of biases found with individuals. Tolcott et al. (1989) studied interacting pairs of intelligence analysts who considered alternative ways in which a scenario might unfold as additional evidence arrived. These experts were given initial information favoring one or another alternative. Their final judgments remained influenced by information received first, rather than the overall weight of evidence. This result exemplifies anchoring and confirmation biases in pairs. A second source of bias in groups emerges from group-level processes (e.g., see review by Kerr et al. 1996). Group discussion is often biased to focus on information shared by groups members (review by Stasser and Titus 2003). In discussions where some information is held by all group members, and other information is held only by some, the information that is shared is more likely to be mentioned and discussed more than the information that is unshared (e.g., Larson et al. 1994). Information held by only one member, even if critical, has fewer chances to inform the collective judgments. This is known as the *common knowledge effect* (Gigone and Hastie 1996). These biases undercut one of the potentially most powerful benefits of groups: bringing together diverse expertise, skills, or knowledge.

A key factor that influences bias in groups is group composition. Diversity of group members is an important predictor of group performance. Schultz-Hart et al. (2000) found that face-to-face groups with heterogeneous initial beliefs about decision alternatives exhibited less confirmation bias than groups with homogeneous initial beliefs. Further, performance in heterogeneous groups was less biased than that of individuals working alone (analyzed as a statisticized group). Moreover, structural diversity in the group can increase the exposure of members to different sources of task information, as shown in a large business field study (Cummings 2004). The chances of better performance increase when more knowledge is shared among diverse group members and with others in the organization (e.g., van Knippenberg and Schippers 2007). However, diversity

may involve additional costs for building common ground among group members and collaborative technology can play a key role in reducing these costs.

2.4. Bias in groups supported by methods or technology

Just as the performance of individuals may be enhanced by technology, the performance of groups may be as well, and developing such technology is a broad goal of our research. To date, two main opportunities have motivated research on technological support for groups: new forms of collaboration and more efficient forms of collaboration. First, computer support augments co-located collaboration and enables distributed collaboration, either synchronous or asynchronous, allowing ways to collaborate not otherwise possible. Second, computer support for groups, whether co-located or distributed, synchronous or asynchronous, can reduce some of the process costs. Group decision support systems and appropriate task structure can reduce process losses and increase process gains in groups relative to unsupported, face-to-face groups (Nunamaker et al. 1991). Reagan-Cirincione (1994) found that computer support and a group facilitator could improve the accuracy of group judgment. The combination of integrated group facilitation, social judgment analysis, and information technology succeeded in neutralizing process costs and enabled group gains. In particular, small groups performed significantly better than their most capable members on two cognitive conflict tasks.

However, as with technological support for individuals, the pattern of success is complicated. An extensive review comparing computer-supported groups with face-to-face groups found that the majority of studies showed no performance difference between supported and unsupported groups (Fjermestad and Hiltz 2001). The effects of medium are presumably moderated by other variables such as group composition, task type, data, and structure of the collaborative process (e.g., Straus and McGrath 1994). Broadly, the research shows that group decision support systems can increase the flexibility of group work (i.e., distributed work) and can improve performance on idea-generation tasks such as brainstorming (e.g., Dennis and Valacich 1993; Gallupe et al. 1991). On the other hand, groupware conditions have shown worse performance in measures such as completion time and efficiency, especially when the task requires close coordination (e.g., Straus and McGrath 1994; Kerr and Tindale 2004). Such mixed results have tempered enthusiasm about technologically supported group work, but have also lead to grounding the research and design efforts on more explicit conceptual models of group process (e.g., Kraut 2003). Later research has tried to address *more specific questions*. Of particular relevance to our work are the studies that investigated how a specific bias might be reduced by a specific tool.

Collaborative software has been developed to address specific group- and individual-level biases. Dennis et al. (1997) developed a group decision support

system targeting the group-level bias of discussion favoring shared information (e.g., Stasser and Titus 2003). They found that their technology benefited or harmed group performance depending on the distribution of members' opinions in the group: technology helped only when there was a minority/majority split of opinion. Hightower and Sayeed (1995) also investigated the problem of biased discussion and found that the benefit of technology interacted with group factors; for example, biased discussion occurred especially in computer-mediated communication with high information load and with more shared information.

Lim and Benbasat (1997) developed groupware tools for reducing the individual-level biases of representativeness and availability and for supporting group decision-making. They studied small decisions-making groups. In a first study, they evaluated the impact of technology support on representativeness bias. The technological support consisted of a computer-mediated communication tool aiding group needs, and a representation tool aiding individual needs. Use of the representation tool reduced the representativeness bias. In another study, Benbasat and Lim (2000) evaluated the effects of technology support on availability bias. In that study an electronic brainstorming tool and electronic mail software were used. Each tool increased the attention paid to items that had low availability, and thus reduced the availability bias.

A few researchers have begun designing and testing tools and methods for reducing confirmation bias in asynchronous collaboration (Cho 2004; Cho and Turoff 2001; Smallman 2008). Cook and Smallman (2007) found that a graphical visualization of evidence reduced the confirmation bias of trainee intelligence analysts who work on individual analysis problems. The pragmatic need to improve performance in complex analysis tasks motivated these researchers to develop a collaborative software environment called JIGSAW (Smallman 2008), which seeks to support analysts at different stages of the collaborative task. In addition to graphical visualizations, the system includes shared work areas for initial clarification of the task and for posting of relevant facts by multiple, distributed analysts. In these ways JIGSAW is similar to our CACHE platform, which was used in the experiment described below.

3. Characterization of analysis tasks

Analysis tasks span very different domains of expertise. An intelligence analyst of a national defense department might assess political stability in a region of the world. A public health analyst might assess epidemiological scenarios about person-to-person transmission of a contagious disease. A business intelligence analyst might assess the future impact of a new product or the viability of a new start-up in a given market. Domains each carry important distinctive properties; for example, intelligence analysis must address active intent to deceive in a way scientific domains need not. However, regardless of the domain, analysis tasks involve formulating judgments about a complex problem on the basis of evidence

and assumptions. Analysis is typically developed and shared collaboratively through written communication, but is constrained by needs for privacy and security as well as the need to share. Further, different collaborating groups and different mixes of collaboration and individual work may be needed at different points in analysis. The challenges in accurately carrying out intelligence analysis have been noted by Trent et al. (2007).

Card (2005) has characterized the analysis process as a set of iterated sub-processes, which often occur in parallel. Specifically, Pirolli et al. (2004) suggest that two interacting loops can be identified in the analysis task: (1) an *information foraging loop* that involves gathering and filtering relevant evidence, and (2) a *sense making loop* (Russell et al. 1993) that involves the development of hypotheses and interpretations that explain or account for the evidence.

We note six core needs or "pain points" in analysis, and how technology might address each need:

- *Large amounts of information*. The quantity of relevant information can overwhelm analysts' limited information processing capacity. Technology can aid managing this information through customized tools for retrieving and filtering relevant data, aggregating and annotating information, organizing and structuring content, tracking the evolution of judgments, and aiding memory.
- *Heterogeneous types of information*. The epistemic status of information differs and hence its role in analysis: direct evidence; interpretation of evidence; hypotheses; judgment of evidence–hypotheses relationship; judgment of relevance and accuracy of evidence; analyst's background knowledge; aims of the analyst; etc. Technology can provide dynamic representations that distinguish and relate different types of information (see, for example, the distinction between evidence in the rows and hypotheses in the columns of an ACH matrix, Figure 2).
- *Unbiased integration of information*. Analysts' expertise lets them perceive meaningful patterns and use higher-order principles to solve problems; however they are still affected by systematic biases. Indeed, the specificity of their expertise can overly narrow their attention to one set of data or perspective. Technology can help analysts compensate for identified biases (e.g., aiding use of unshared information) or alert analysts about threats to a final unbiased decision (e.g., flagging the tendency to confirm initial beliefs).
- *Probabilistic, multi-stage inference*. Analysts make chains of probabilistic inferences in which a prior inference is input to the next inferential stage (Johnson and Halpin 1974). Representation tools have a critical role in reducing such error at the individual level (Lim and Benbasat 1997). Collaborative systems can also reduce error by supporting group-level and community-level checks. Collaborative technology can help analysts to check one another, to make the problem decomposition visible, and to track and update dependencies among parts. Technology can increase the

collective awareness of inconsistencies and promote active, group self-correction (e.g., balancing the contributions from different members during the analysis process).

- *Justifications and information about provenance*. Meta-information about the source and reliability of information (provenance) should be available for inspection when developing or defending judgments. Further the status of information and of source can change, for example, when the provider of a piece of evidence discredits it. Collaborative technology could dynamically track provenance and support use of provenance in collaborative reasoning and evaluating arguments.
- *Multiple, dynamic modes of work*. Even within a given task, an analyst may shift among tightly coupled discussion, loosely coupled information sharing, and working independently of others. Such transitions add to the overhead of coordination and can cause repeated losses of context, particularly when transitions force the analyst to switch among tools (e.g., personal notepad, phone, email, shared database). Collaborative technology can provide an integrated task environment that preserves useful contextual information as analysts move across the various work modes: multiple tools in an integrated environment can allow analysts to dynamically select the set of tools that best support the current work mode.

4. Goal, hypotheses, and design of experiment

The goal of this experiment was to investigate the impact of group composition on judgment bias and on information coverage in small distributed groups supported by collaborative software. *Our general hypothesis was that the members of groups with diverse initial beliefs about the analysis would collectively explore more of the evidence space and better counter cognitive biases*. Schultz-Hart et al. (2000) found that groups homogeneous with respect to the members' initial beliefs (i.e., prior to group discussion) are more biased than heterogeneous groups. Note that this study investigated face-to-face teams working on a simple decision-making task and measured bias only during the selection of the evidence.

We studied bias in the context of computer-supported collaboration. Our experimental task involved loosely rather than tightly coupled collaboration (see work coupling in Neale et al. 2004). The collaborators shared and discussed evidence and analyses in developing their individual product. The task product was the set of individual analyses rather than a single group product or decision. The interdependencies to be managed pertained to shared resources and process: the members analyzed the same body of evidence, *synchronously* developed individual analyses, *iteratively* accounted for the collaborators' analyses and updated their beliefs as new evidence was analyzed and discussed. Our

collaborative technology supported individual work and sharing, distribution, and discussion of this work across the distributed group.

We extended prior research in several ways. First and most importantly we developed an environment for collaborative analysis, CACHE-A, described below, and used this as our test bed to investigate computer-supported, rather than face-to-face collaboration. Thus, in terms of the quadrants in Figure 1, our study investigates collaboration with small, distributed groups and rich technology support (note that the general CACHE architecture is designed to support a distributed community). Second, we used a realistic, moderately complex analysis task, unfamiliar to our participants. We developed an experimental analysis task containing multiple hypotheses and many conflicting pieces of evidence; this invites complex reasoning about disconfirmation, negative evidence, and multiple alternatives, typical of many naturally occurring decisions. Because participants were unfamiliar with the particular problem scenario and we had a flexible evidence pool, we could induce and control alternative biases in our participants. This task was designed both to provide experimental control and to model important and under-investigated aspects of intelligence analysis tasks. Third, we studied bias (and bias change) not only at the level of evidence item selection but also at the level of summative judgments (i.e., beliefs about hypotheses); we also collected information about the decision making process, including evidence selection, assessment, and integration. In addition to testing the effects of our manipulation, this study also provided our first opportunity for user feedback and assessment of the prototype.

We induced an initial bias in the beliefs of individual participants (i.e., Initially Preferred Hypothesis). On this basis we manipulated the group composition to form groups with heterogeneous or homogeneous members' initial beliefs: this is the Initial Belief factor in the experimental design. Consistently with prior research, we included groups of non-interacting individuals – called solo or nominal groups. We asked how an individual's performance was affected by the composition of their group and we predicted that:

- *Heterogeneous groups would show less confirmation bias than Homogeneous groups*. Because CACHE-A supports sharing and comparing information among participants, the heterogeneous groups should mitigate cognitive biases by exposure to (1) more and more balanced evidence and (2) a more diverse set of judgments.
- *Heterogeneous groups would show no net process loss relative to the Solo/ Nominal Group*. Because CACHE-A mitigates the process costs, heterogeneous group should show equivalent or better performance.

4.1. Factors

Group Condition was a between-subjects factor with three levels: Homogeneous, Heterogeneous, and Solo (or Nominal) groups of three individuals.

Initial Belief was a between-subjects three-level counterbalancing factor, orthogonal to Group Condition. Table 1 shows the assignment of the participants to Group Condition and Initial Belief (Initially Preferred Hypothesis: H1, H2, *or* H3). In the Heterogeneous condition, the three group members each had a different Initial Belief or preferred hypothesis (i.e., H1, H2, H3). In the Homogeneous condition, the three group members all had the same Initial Belief (e.g., H1, H1, H1). In the Solo, or Nominal, condition individuals did not interact; for purposes of experiment running, sets of three individuals were scheduled together and assigned heterogeneous initial beliefs. However, these individuals could be equally assigned post hoc to "groups" with homogeneous or heterogeneous Initial Belief (see statisticized groups in Schultz-Hart et al. 2000).

Judged Hypothesis was a within-subjects, counterbalancing, three-level factor. Participants entered responses for each of the three hypotheses (e.g., in the columns of an ACH matrix: H1=friction, H2=over-ram, *and* H3=suicide). Judged Hypothesis identifies which hypothesis.

Block was a within-subjects, four-level factor. The evidence was presented in four blocks. After each block the participants rated their belief (i.e., degree of confidence) in the three hypotheses.

4.2. Response variables

Bias Measures: (1) Rated Belief in the preferred hypotheses (in the CACHE matrix) and its change over blocks of evidence were our most general measures of bias. (2) Additional bias measures were derived from components of the final ACH matrix: the amount of positive evidence included in the matrix (Inclusion),

Table 1 Assignment of participants to initially preferred hypothesis and condition.

Preferred hypotheses/condition	Group number	Members' preferred hypotheses
Heterogeneous	Group 1	H1 H2 H3
	Group 2	H1 H2 H3
	Group 3	H1 H2 H3
Homogeneous	Group 4	H1 H1 H1
	Group 5	H2 H2 H2
	Group 6	H3 H3 H3
Solo	Group 7	H1 H1 H1
	Group 8	H2 H2 H2
	Group 9	H3 H3 H3
Total	9 groups	27 participants

H1, H2, H3 in the rightmost column indicate members' initial bias toward Hypothesis 1 (friction), Hypothesis 2 (over-ram), or Hypothesis 3 (suicide), respectively. The rows indicate conditions and groups by condition. 27 participants are arranged into nine groups. This assignment balances the three hypotheses across the three conditions. The groups in the Solo condition are formed by triples of non-interacting individuals.

the rated importance of positive evidence (Importance), and the rated relation between the evidence items and hypotheses (Weight). The measures were derived by computing the difference between values for the preferred hypothesis versus for the average for the alternative hypotheses. These measures are the Inclusion Score, Importance Score, and Weight Score. 3) In addition, we looked at a system-integrated measure of strength of evidence in the matrix (Strength).

Coverage: amount of relevant information analyzed, which was measured as number of relevant evidence items that were included in the final matrix, independent of bias.

System and task evaluation: We also evaluated the CACHE-A system and the task through a usability questionnaire and a brief interview. Logging data and qualitative comments were also collected for a broader picture of the activity.

5. Method

5.1. Participants and setting

Twenty-seven participants, primarily university students, were assigned to three-member same-gender groups (nine groups). Two additional groups were run but excluded from the data analysis because of procedural or technical problems. Of the 27 participants, about one third (10 of 27) were females. The average age was 25.2; two participants were older than 40 years. The participants were not experienced intelligence analysts and they were unfamiliar with the domain. This facilitated investigating the role of bias in analysis: (1) Judgment bias effects on novices are sometimes larger than on experts, and (2) Lack of opinion about the case prior to the experiment facilitates uniform manipulation of bias.

Participants in a given group trained together and thus gained some familiarity with one another. Following training, the members of each group were seated at workstations located in separate rooms. They could talk to the experimenter through a chat tool. The members of interacting groups could share information with their partners using the chat tool and the collaborative components of CACHE-A (described below).

5.2. CSCW tools

5.2.1. *CACHE architecture*

CACHE (Shrager 2005) is a software system that simultaneously supports the work of individual analysts and various forms of collaboration. CACHE allows individuals to efficiently share information developed either inside or outside of the intelligence community, such as signal and operational intelligence or open source web-based and news reports. The CACHE architecture is designed to support intelligence analysis at the level of a distributed community, to which

individual analysts belong and contribute. Information provided by one community member sets up priors which others may build further knowledge upon.

The CACHE software supports analysis by easing the six "pain points" characterizing analysis (Section 3): (1) search tools for handling large amounts of information and visual feedback on relevance of the search results; (2) representations that relate different information types including Evidence-by-Hypotheses matrices, an evidence pool, and support for annotation (i.e., interpretations); (3) automatic integration of analyst-supplied estimates in the decision matrix: in the matrix the system displays an up-to-date summative estimate of the current support for each hypothesis; (4) weighted ratings to convey uncertainty (when used by a community, the system makes reasoning chains explicitly represented and available); (5) provenance maintenance, available for inspection upon demand (in the experiment, the system used both provenance and reasoning chains to enable real-time up-to-date incremental reasoning); and (6) a stable work environment supporting varied forms of collaboration, including shared representations developed by analysts in individual workspaces. The CACHE architecture integrates several technologies for supporting analysis. For example, in addition to exact-match searching, its natural language technologies provide semantically driven search and ranking of relevance of the results. This architecture provided us with a useful testing environment for investigating collaborative analysis. The experiment used a simplified prototype of CACHE, CACHE-A, developed specifically for this study.

5.2.2. *CACHE-A and chat tool*

CACHE-A includes a subset of the full CACHE functionality, combined with a control system that ran the experiment. The control system handles most aspects of running the experiment, including providing each user access to the right collaboration conditions block by block. For example, for interacting groups in the experiment, each of the blocks 2, 3, and 4 began with five minutes of individual work, then transitioned to collaborative work, and ended with a request to update current beliefs based on the evidence so far. CACHE-A provides the pool of evidence appropriate for each user, block-by-block. It also logs operations by the participants, the experimenter, and the system itself.

CACHE-A provides a graphical interface for the analysis task (Figures 3 and 4). Three tools support individual work: the individual's ACH matrix, the search tool, and the evidence viewer (Figure 4, bottom windows). Three other tools support collaborative work: the ticker, two views of partners' ACH matrices and the chat tool (Figure 4, top windows).

The central component of the system is the user's ACH matrix. Like ACH0, CACHE-A supports the ACH method (Heuer 1999) by permitting each analyst to enter pieces of evidence into her/his own ACH matrix. Figure 2 shows an

example of an ACH matrix. Rows represent evidence, columns represent hypotheses, and the row×column cells represent the relation between the corresponding piece of evidence and hypothesis. All the pieces of evidence selected by a given analyst are listed as the rows of his or her matrix. The first column contains the heading of each piece of evidence and the second column allows the analyst to rate importance of that piece of evidence. The remaining columns of the matrix represent all the hypotheses considered by the analyst (e.g., the three hypotheses in Figure 2), and remain in a fixed order. The values users assign to the row×column relational cells range from strongly consistent (CC), consistent (C), neutral or not applicable (N/A), inconsistent (I), to very inconsistent (II). The color of the cells is green for the most informative values, directing attention to strong disconfirmation. The pieces of evidence for or against the hypotheses can be added or removed to the ACH matrix, but the hypotheses cannot be modified. At the bottom of the matrix, in purple, the Belief row allows the user to enter their degree of belief in each of the three hypotheses. Belief across all hypotheses must total 100, and the user is prompted if other combinations of values are tried. While the Belief row is user generated, the Support row provides a system-generated summative measure of support, integrating the information the user provided.

The example in Figure 2 shows the three hypotheses used for all conditions, "Over-ram," "Friction," and "Hartwig" (i.e., suicide). In the first row user ALPHA has entered a piece of Navy test evidence. ALPHA has rated this piece of evidence as highly important and as strongly inconsistent with the Friction hypothesis (in green), and has set consistency relations with the other two hypotheses as well. Across the whole matrix, the current values ALPHA last assigned to Belief give the highest rating (of 50) to the Friction Hypothesis. As indicated in the Support row, currently it is the Overam, not the Friction Hypothesis, that is least disconfirmed according to the user's ratings; the user may chose to align summary belief with current evidence when he or she next updates belief.

The search tool (Figure 3, bottom left; Appendix: Figure 8) works much like Google's advanced search interface and its window opens from the link at the bottom of the ACH matrix. A search box and a search button enable keyboard-based queries. In addition to the keyword, the analyst can specify options in (1) scope of search (over evidence or interpretations), (2) display (as concise or full listing), and (3) search method (e.g., exact keyword match or semantic similarity). The experiment used default settings of searching evidence, concise display, and exact-match search method. The search results are listed with the most relevant at the top and in larger font.

The evidence viewer (Figure 4, bottom right; Appendix: Figure 9) is used to read detailed information about a specific piece of evidence. A piece of evidence is accessed and displayed through its interpretation, which functions like a header or title. Each item is given one initial interpretation, but users may add more. Clicking on an interpretation displays the longer text making up the piece of

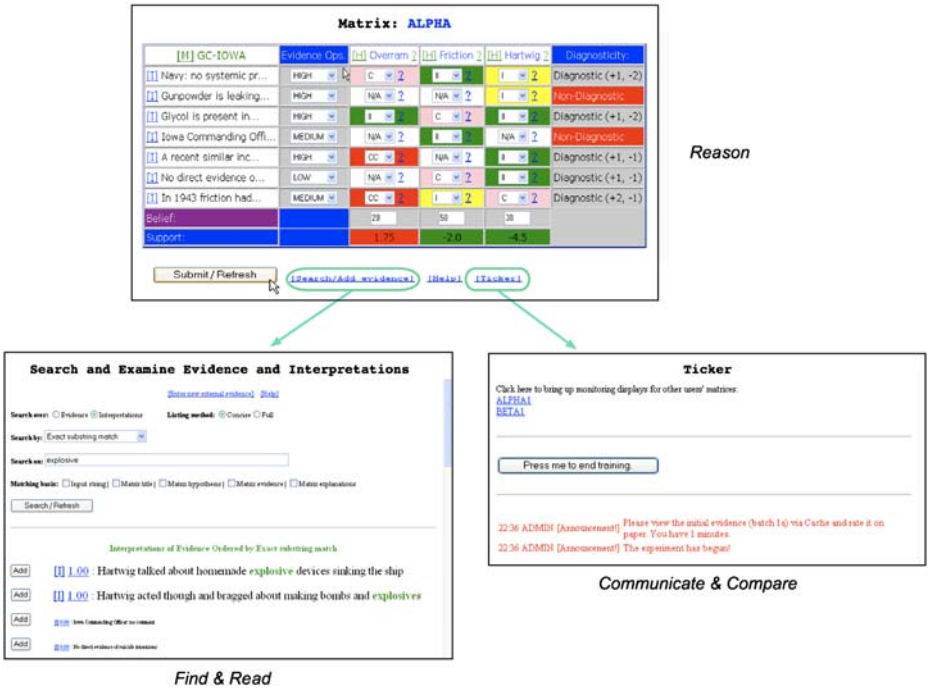

Figure 3 CACHE-A: three windows for three functions. The *analyst's ACH decision matrix*, the *search tool* for finding evidence, and the *ticker*, which reports on system events and between-matrices differences between partners' analyses and the user's own analysis.

evidence. Users add the evidence title, or interpretation, to their matrix. The evidence header retains its link to the evidence, wherever the header appears, allowing the user to click open the evidence window from multiple access points. For example, by clicking on the link of the header listed in the ACH matrix (see [I] links in Figure 2) or in the search tool (Appendix: Figure 8) the analyst will open a separate window for each piece of evidence. Multiple evidence viewers can be open at once, allowing comparison of multiple pieces of evidence, but requiring window management.

Turning to the collaborative tools (Figure 4, top windows), the ticker (Appendix: Figure 10) is a notification system that conveys automatically generated information to participants. It displays system information that guides the course of the experiment and automatically captures information updating one partner on potentially relevant actions of another. The analyst will be updated concerning the partners' newly added evidence and newly made judgments about particular pieces of evidence (i.e., ratings). The ticker of a given analyst, Alpha, will report when her/his collaborators, Beta or Gamma, add a piece of evidence to their individual matrices that is not already in Alpha's matrix. Moreover, if Alpha and Beta both have entered a given piece of evidence into their individual matrices already, then if Beta enters a different judgment about that piece of

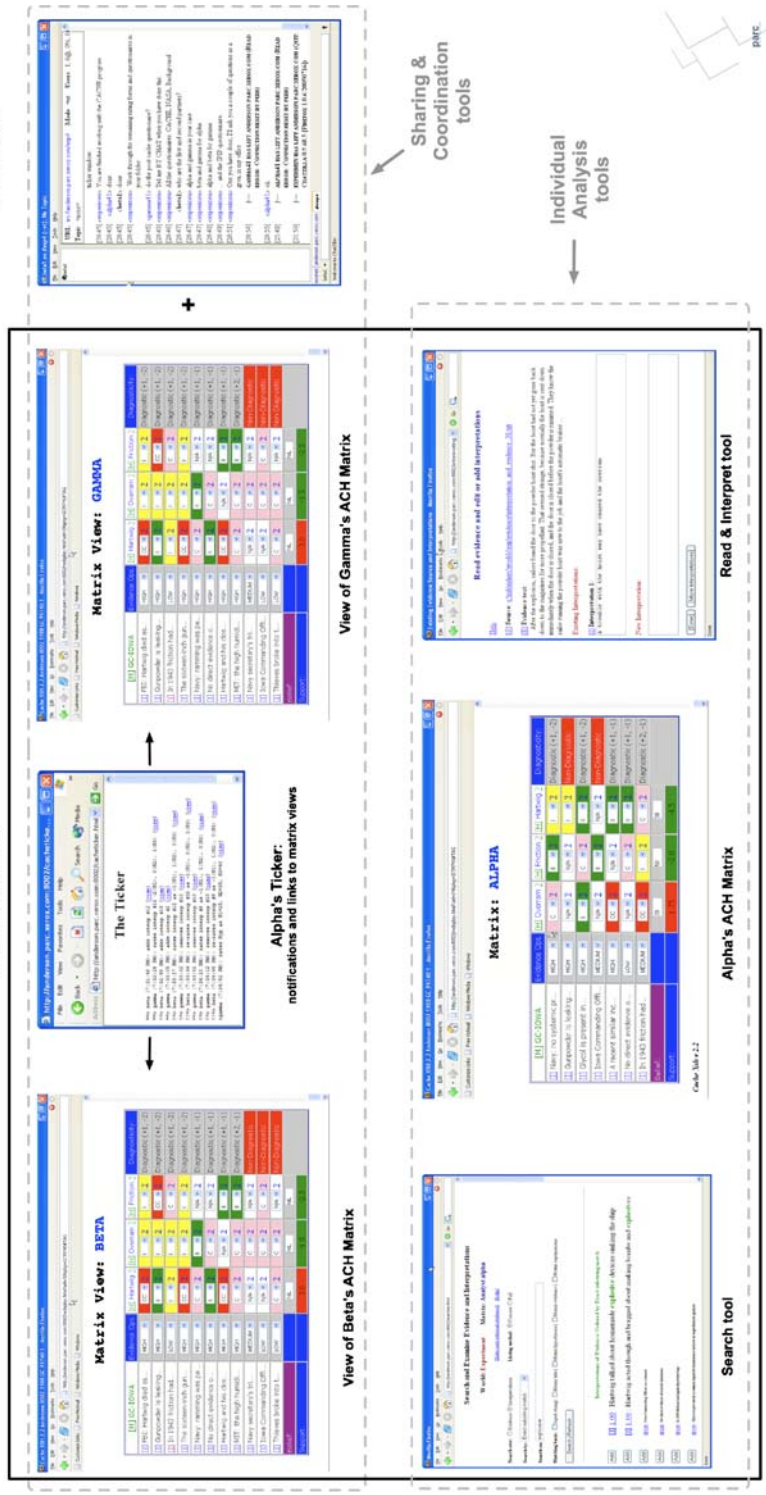

Figure 4 CACHE-A Workspace for one team member (Alpha). Search tool, ACH matrix, and Evidence viewer (*Read & Interpret*) are used for individual analysis (*bottom*). Read-only views of partners' matrices, Ticker, and Chat tool support collaboration (*top*).

evidence (e.g. Beta judges it very inconsistent whereas Alpha judges it neutral), a one-line notification of this difference will be posted in Alpha's ticker. Each posted notification ticker includes a link. By clicking on the link, the user can open and read the details of the piece of evidence in the evidence viewer. Thus, the ticker keeps group members aware of differences in their developing analyses.

Two read-only views of partners' ACH matrices (Appendix: Figure 10) also provide a channel for collaboration, showing the partner's evolving analysis. During collaborative work, each partner's matrix can be accessed through a link in the Ticker. These views allow *monitoring* in real time (but do not allow changing) the status of each partner's ACH matrix. An analyst may learn about a new piece of evidence and *borrow* it through a partner's matrix as well as through a ticker message. By clicking on the evidence header one user can gain access to information in the partner's evidence pool, whether or not that evidence was already available in the user's own pool. These two methods – ticker and viewing partner matris – complement each other, one highlighting a recent change and the other showing the whole context.

Finally, an off-the-shelf chat tool allows each participant to communicate with partners during the collaboration phase. It allows participants to talk about any topic they believe valuable to discuss with their partners.

During the experiment, particularly in the collaborative mode, participants had to manage multiple windows at once, such as multiple matrix views as well as evidence viewers. The chat and the ticker tools were also used by the experimenter to distribute synchronously scripted instructions to the participants, during the experiment.

5.3. Materials

Participants used the CACHE-A system to conduct a moderately complex analysis task, designed to model a real analysis problem, the USS Iowa case.

In 1989 an explosion occurred on the battleship USS Iowa in one of the battleship turrets, taking a complete toll of all forty-seven crewmen in that turret. Months later, the results of the investigations by over fifty experts from several laboratories identified three possible causes of the explosion with evidence supporting each of the three hypotheses. The cause favored by the Navy was human error resulting in *over-ramming* of the powder bags into the gun; that favored by Sandia National Labs was *friction*-produced static electricity igniting the powder; while the FBI concluded that an incendiary device had been used in attempted *suicide*.

Cheikes et al. (2004) used this case as a task for studying bias with individual analysts. We redesigned and extended their task materials. Our task was designed for measuring judgment bias in computer-supported collaborative analysis. It differed from the version used for individual analysts in significant ways. First, we used an extended evidence pool after pre-validating the items with three independent judges (i.e., 80 evidence items). Second, we added an interpretation

Table 2 Stimuli: evidence items by block and hypothesis.

Block/hypothesis	Block 1	Block 2	Block 3	Block 4	Total items
Preferred hypothesis	4 (all P)[a]	6 (5P, 1N)	3 (1P, 2N)	3 (1P, 2N)	16 (11P, 5N)
Alternative Hyp.1	4 (all P)	3 (2P, 1N)	4 (2P, 2N)	5 (3P, 2N)	16 (11P, 5N)
Alternative Hyp.2	4 (all P)	3 (2P, 1N)	4 (2P, 2N)	5 (3P, 2N)	16 (11P, 5N)
Fillers (Neutral)	8	8	9	7	32
Total items	20	20	20	20	80

The columns indicate blocks and the rows hypotheses, to which the items are related. Positive items (P) confirm and negative items (N) disconfirm one hypothesis. Fillers are neutral (nondiagnostic). The first five items analyzed individually in Block 1 include four positive items pro preferred hypothesis ([a]) and one filler. The distribution of the items is used to induce bias.

to each piece of evidence in the pool. Third, we introduced three analyst roles (Navy, Sandia, and FBI expert), where each role modeled an expert with distinct knowledge and preferences on hypotheses. Finally, we controlled the composition of the evidence pool by balancing the evidence in relation to the three hypotheses and the three analyst roles (e.g. an equivalent number of items was tagged with Navy, Sandia, or FBI as their source).

The entire "case book" for the analysis included background material about the Iowa case, the three hypotheses, and 80 evidence items (33 positive items, with three sets of 11 items supporting each of the hypotheses; 15 negative items, with three sets of five items disconfirming each hypothesis; and 32 neutral or nondiagnostic fillers). Negative or "disconfirming" evidence here means directly and relevantly arguing against one of the hypothesis, not simply supporting an alternative. Each evidence item consisted of the base text description of evidence and a short one-sentence interpretation, acting like a title or headline. The title clearly indicated whether the evidence item supported, opposed, or was neutral with respect to the relevant hypothesis. Overall, the evidence covered many topics and contained conflicting expert testimony connected to different sources, including consulting agencies (e.g. MIT) as well as the main agencies (Navy, Sandia, and FBI).

The 80 pieces of evidence were all viewed by all participants and were partitioned into 4 ordered blocks, as shown in Table 2. The order of presentation of the evidence was controlled and differed depending on the Initial Preferred Hypothesis assigned to the participant (see description below). Before the experiment, three judges had rated the consistency relationship of a larger corpus of evidence items in relation to the three hypotheses. The ratings were used to select 80 unambiguous items and to balance the support across the item set, for the three hypotheses. Overall, the support of the final set of 80 items for the three hypotheses was quite even. An ideal unbiased analyst *after considering all the evidence* was expected to update her/his beliefs in the three hypotheses until they would become approximately equal across the three hypotheses (i.e., 33% probability for each hypothesis).

Table 3 Procedure by group condition.

Phases/ condition	Training (50 min)	Block 1 (25 min)	Blocks 2, 3, 4 (20 min each, 60 total)	Questionnaire, interval (16 min)
Hetero	1. Group (35 min): ACH and system; Practice in the CACHE-A system	1. Indiv. (5 min): Read 5 items; Rate 5 interpretat. (bias check)	1. Indiv. (5 min): Analyze new block items, add to matrix, rate; Rate belief in hypotheses	1. Indiv. (8 min): Post-task survey on system and task
	2. Indiv. (15 min): Learn role, task, and case; Background survey	2. Indiv. (20 min): Analyze block 1, add to matrix, rate; Rate belief in hypotheses (bias check)	2. Group (15 min): View partners' matrices, ticker, check new items, discuss in chat; Rate belief in hypotheses (bias check)	2. Group (8 min): Interview on task, system, process; Debriefing
Homo	As above	As above	As above	As above
Solo	As above	As above	1. Indiv. (5 min): (see Indiv. above) 2. Indiv. (15 min): (see Indiv. above)	As above

The columns indicate the phases and the rows the condition. Training and Question Phases included group and individual work in all conditions. Blocks 2, 3, and 4 included (distributed) group work only for the Homogeneous and Heterogeneous conditions.

The experiment was designed to induce in each participant an initial bias favoring one of the three hypotheses. We refer to the hypothesis toward which a participant was initially biased as the Initially Preferred Hypothesis, or Initial Belief. Such initial bias was induced by two means. First, each participant was assigned a distinct analyst role (Navy, Sandia, or FBI expert). The role primed the participant with a preference for evidence from her/his agency and a particular type of etiological argument (human factors, mechanical, or motivational). Second, the order of presentation of the evidence was manipulated in several ways within and across blocks: (a) in the first block, the first five items analyzed included four positive items strongly in favor of the preferred hypothesis and one neutral item – this primed the analyst to "anchor" on this hypothesis; (b) across the four blocks, the participant analyzed more positive evidence in support of the preferred hypothesis in the earlier blocks than positive evidence for an alternative hypothesis; and (c) within each block the evidence was ordered so that stronger confirmatory evidence for the preferred hypothesis was presented earlier.

In summary, we induced bias at the beginning of the task, but by the end of the task each participant had viewed the identical, and balanced, evidence set. The experiment design is similar to those used in prior studies of bias with experts

(e.g., Cheikes et al. 2004; Tolcott et al. 1989) or non-experts (e.g., Schultz-Hart et al. 2000).

5.4. Procedure

For all conditions, the procedure consisted of training, the core analysis task, and final questionnaires and interview. Table 3 summarizes the phases of the procedure in the three conditions. The core analysis task took place in four blocks. The conditions differed in whether the task included collaboration.

Participants were trained in groups of three, in a meeting room for about a half hour. They received a preliminary description of the study and signed informed consent. They were introduced to and practiced with the ACH method and the components of the CACHE-A system. They moved to workstations located in three separate rooms where they had about 15 min to read the descriptions of their role (Navy, Sandia, or FBI expert), background information about the case, and the three hypotheses.

The main analysis task lasted about one hour and half and was structured in four blocks, lasting about 20-min each. In each of the four blocks, the CACHE-A server presented each group member with the appropriate block of 20 evidence items. Throughout the analysis task, participants used the CACHE-A tools to find and read evidence that they considered relevant, and gradually added the evidence selected to their ACH matrix. Any time, after adding a new item of evidence to the matrix, the participant could specify the importance of that item and its relationships to any of the three predefined hypotheses in the matrix (consistent, inconsistent, or neutral; see Appendix: Figure 7). Participants could update their belief in the hypotheses at any time, but were requested to do so at the end of each block, assigning 1–100 belief ratings at the bottom of the matrix. Belief judgments were constrained to sum to 100 across the three hypotheses (see Figure 2).

During the first block all the participants worked alone. For each of the remaining three blocks the members of interacting groups (Homogeneous and Heterogeneous groups), were able to collaborate with one another through the ticker, the views of the partners' ACH matrices, and the chat tool. At the beginning of each block, they worked individually for five minutes, and then they could interact with their partners for about 15 min. For the collaborative part, the participants were invited to share information ("use the information from your teammates to help your own analysis, whether or not their opinions are similar to your own") and discuss differences via chat. The members of Solo (nominal) groups worked individually for the entire duration of the analysis task. The time they had to work on each block was the same as in the two interacting conditions; that is, collaborating individuals had no additional time for communication during analysis, nor did they have additional time for training on the collaborative tools. After finishing the four blocks of the analysis task, participants answered a usability questionnaire, took part in a brief interview, and finally were debriefed.

5.5. Illustrative scenario: a team working in CACHE-A

Three participants are working as analysts on the USS Iowa case. Analyst "Bertram", who plays the role of the Sandia analyst, suspects that the explosion was caused by friction producing a spark. At the start of the analysis task, Bertram scans the two monitors in his office, checking his empty ACH matrix, ticker, and chat windows. He recalls that his own matrix and the ticker should always remain open. Then the ticker reports the availability of evidence. Bertram searches the evidence for "friction" and reads over the interpretations presented, summarizing pieces of evidence. He begins to open and read evidence. He clicks the ADD button and watches the item appear in his matrix. He sets the value of the evidence as moderately important and changes the cell indicating its relation to the Friction hypothesis from the default Neutral value to Strongly Consistent. This item doesn't seem very relevant to the other hypotheses so he leaves those cell values at neutral. Presently he reads the ticker message announcing that the collaborative part of the block has begun. Bertram finishes up a couple more pieces of evidence and then opens "Alphred's" matrix. Alphred has a lot of evidence entered but not rated. However, Bertram notices that Alphred has a piece of evidence marked as strongly inconsistent with the friction hypothesis, so he opens and reads this. "Gamba" has posted a question asking what the others think about a particular piece of evidence. The three analysts continue working with a mix of individual and shared activity through the remainder of the block. The cycle of individual and group work continues across blocks. The group chats more toward the end of the last block, asking about the others' assessments.

6. Results

6.1. Overview

Across all bias measures the pattern of results were the same, though it varied in whether or not it reached significance: groups in the homogeneous condition always tended to be more biased than those in the heterogeneous and solo conditions, which were similar to one another. We report bias effects on two types of measures in Sections 6.2 and 6.3. (1) The primary measure of bias and debiasing was the overall judged Belief in the preferred hypothesis, including changes in belief over blocks. This is the value that users entered in their matrix in the Belief row. (2) Additional cells, from the final matrix, provided finer grained measures that can suggest what aspect of judgment contributed most to bias: amount of positive evidence included (Inclusion) and assessment of that evidence (Importance and Weight). For these measures we assessed bias using a difference score, comparing values for the preferred versus alternative hypotheses. We also used the system-generated, unbiased integration of user-provided values (Strength).

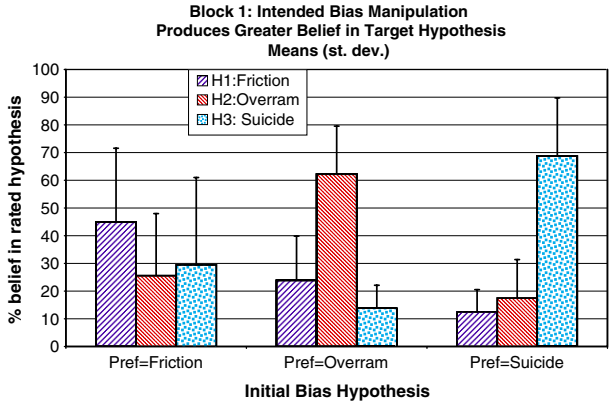

Figure 5 Initial Belief (after *Block 1*). Successful bias across conditions and Initially Preferred Hypotheses.

Independently of questions of bias, we looked for condition effects in the amount and type of evidence included in the final matrix (Section 6.4). Finally we take a broader perspective about using the CACHE-A tool and report some on-line tracking data and highlights from the usability questionnaire and debriefing (Section 6.5 to 6.6). Some logging data collected during the course of the experiment was missing: where a measure was not captured for all users, the actual *N* is noted. Note that this exploratory study used a small number of groups and individuals, so our statistical sensitivity is modest and best suited to detecting robust effects.

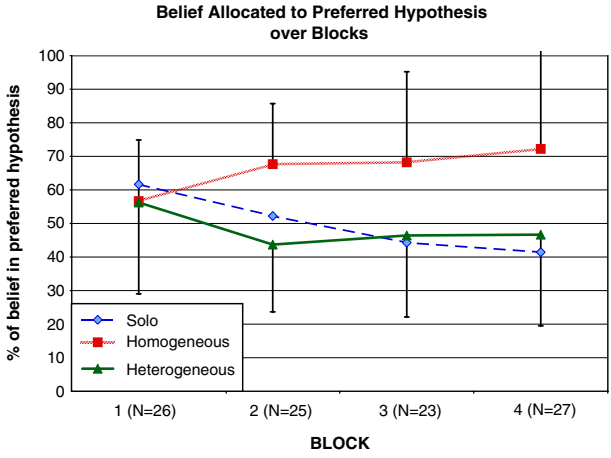

Figure 6 Belief (%) in the Initially Preferred Hypothesis at the end of each block. Standard Deviation for Homogeneous and Heterogeneous Conditions (*vertical bars*). *N* Number of participants with data available for each block (on the *x*-axis).

Table 4 Bias measures by group condition.

Bias measure	Condition (averages)			
	Hetero	Homo	Solo	Total average
(a) Inclusion difference score={(# pieces of positive evidence items included for the preferred hypothesis)−(# for average of the alternatives)}	1.61	2.17	1.17	1.65*
(b) Importance difference score={(ave. rated importance of positive evidence items for the preferred hypothesis)−(importance for average of the alternatives)}	1.07	1.51	1.17	1.25*·**
(c) Weight difference score in consistent/inconsistent relation={(ave. rated consistency of positive evidence items for the preferred hypothesis)−(value for average of the alternatives)}	3.06	8.22	1.72	4.33*
(d) Strength={CACHE-computed summative score of strength of support for preferred hypothesis}	3.67	6.75	−1.33	2.88*·**
(e) Belief={Overall rating of belief in the preferred hypothesis}	46.67	72.22	41.44	53.44*·**

Difference scores (rows a, b, c) aggregate judgments expressed on the evidence items. Absence of bias for these scores corresponds to zero: i.e., large values indicate greater bias. The Strength and Belief scores indicate bias at the level of the entire ACH matrix (i.e., belief in preferred hypotheses).
*Significant bias overall, $p>0.05$
**Significant EFFECT of Group CONDITION: condition difference in bias, $p>0.05$

6.2. Belief bias: initial, final, and change

To get a clean measure of bias change, we needed to induce strong, analogous initial bias across conditions. We succeeded as shown in Figure 5. We produced our intended initial bias toward the preferred hypothesis, targeted for the individual user, $F(4, 34)=8.49$, $p<0.001$ (Condition×Initial Bias, repeated measures), without Condition or interaction effects ($F<1$). Overall, 58% of Initial Belief was committed to the preferred hypothesis, greater than the uniform distribution of 33%, consistently across conditions (Heterogeneous Condition mean=56, Homogenous Condition=57, Solo=62).

However, belief changed very differently across conditions, as shown in Figure 6. The Belief in Preferred Hypothesis in the Heterogeneous and Solo Conditions dropped from roughly 60% to 43%, showing strong debaising effects from exposure to balancing evidence. In contrast, the Homogeneous Condition showed polarization of belief, from 57% to 72%. Group Condition significantly affected the amount of Bias at Block 4 (ANOVA $F=3.92$, $p=0.034$) and was marginal or significant at Block 2 ($F=3.8$, $p=0.041$) and Block 3 ($F=2.93$, $p=0.076$). A MANOVA produced analogous findings. Repeated measures were not used due to missing data in intermediate blocks, as indicated in Figure 6. Table 4 Line e shows belief from the final matrix.

6.3. Bias and components of the final matrix

Users' final beliefs showed bias and effect of condition, so we looked at other measures from the final matrix that might provide finer grained information about these effects. The final matrix captures a great deal of information about what the user did in addition to their final 0–100% belief rating, including what evidence the user selected for the matrix, how the user rated the importance of a piece of evidence, and how the user judged the relation between any particular piece of evidence and a particular hypothesis. In addition, the system displays at the bottom of the matrix a summative score of information in the matrix cells for each hypothesis. These judgments build on each other: evidence cannot be rated for its importance or relation until it is included in the matrix, and the summative score computed by the system uses the data previously entered in the cells. We consider component processes in reaching a final belief by analyzing these measures. First, we looked for bias and effect of condition on the selection of evidence included in the matrix. Second, we looked for bias and effect of condition on assessing evidence, both evidence importance and weight of the evidence as consistent or inconsistent with a hypothesis. Third we looked for the effects on the system-generated summative score of information in the cells; this approach can identify bias in components separately from possible bias in integration of those components, which are combined in the user-entered belief ratings.

Inclusion of Evidence. Users were biased in the evidence they included, but we did not find significant differences among conditions. Users were significantly *more* likely to include *positive* evidence (supporting an hypothesis) for their favored (79%) than for the alternative (64%) hypotheses, a difference of 1.65 (of 11) more pieces of evidence per hypothesis, $t=(26)=4.1$, $p<0.01$. Table 4 *Line a* shows inclusion of positive evidence as the difference for the preferred hypothesis and the average for the alternatives. A repeated measures analysis of the effect of Preferred vs. Alternative hypothesis (with Condition, Initial Bias, and all interaction terms) compared number of pieces of positive evidence for the favored hypothesis to the average number for the two alternative hypotheses, confirming the significant bias effect of Preferred vs Alternative Hypothesis $F(1, 18)=17.18$, $p=0.001$, eta2=0.49, but found no effect of Condition $F(2, 18)<1$, eta2=0.056; of Initial Bias $F(2, 18)=$ 1.1 eta2=0.109; or interactions (all Condition interaction Fs<2, ps>0.15, etas<0.3). Analogously, users were significantly *less* likely to include *negative* evidence (refuting an hypothesis) for favored (64%, 3.2 of 5) than for alternative (78%, 3.9 of 5) hypotheses, $t(26)=-2.714$, $p=0.012$; the analogous repeated measures analysis found effects of Preferred vs Alternative Hypothesis $F(1, 18)=6.86$, $p=0.017$, eta2 (0.276); but no effect of Condition $F(2, 18)<1$, eta2=0.056; of Initial Bias $F(2, 18)=$ 1.1 eta2=0.109; or of any interactions $F<1.2$, eta2<0.2. The inclusion rates and bias level for negative and positive evidence are similar to each other.

Assessment of Evidence. Once evidence is entered, it may be rated for importance and its relation to any hypothesis noted. Users rated the importance of positive evidence supporting the favored hypothesis more highly than the importance of positive evidence supporting the alternative hypotheses, measured with a difference score in Table 4 line b. This difference score is significantly greater than zero, $t(26)=4.687$, $p<0.001$, showing bias. Degree of bias differs significantly among conditions, Group Condition, $F(2, 18)=5.999$, $p=0.01$,eta= 0.40; The homogeneous, and solo, conditions showed significant bias ($t(8)=6.416$, $p<0.001$ and $t(8)=2.485$, $p=0.038$, respectively) but the heterogeneous condition did not $t(8)=1.037$, $p=0.33$. There was also a significant effect of Initial Bias $F(2, 18)=4.418$, $p=0.27$,eta2=0.329.

Users also showed bias in how evidence was related to hypotheses, as expressed in the five-point scale ranging from very consistent to very inconsistent. Table 4 line c shows the difference scores for Weight, with average consistency of evidence rated higher for the favored rather than alternative hypotheses, $t(1, 26)=2.497$, $p=0.019$). However, no analysis showed significant effects of Group Condition, or of Initial Bias.

Integrating evidence. To assess the implication of the information in the evidence × hypothesis cells, separately from any effects of biased integration, we looked at the strength index provided by CACHE-A. CACHE-A sums the evidence weighed by degree of support and importance ratings (scaled around zero) to provide a strength index, which was displayed to the user under each hypothesis. This provides an unbiased integration method of all the matrix cell information, hence bias here would only reflect the cumulative effect of bias in component information, not biased integration as might occur in forming final belief. Overall, the strength index for the favored hypothesis was greater than zero, showing significant bias, $t(1, 25)=2.243$, $p=0.034$, and carried primarily by the Homogeneous condition. The strength index for the favored hypothesis differed across conditions, as shown in Line d Table 4, ANOVA $F(2, 23)=4.16$, $p=0.029$, eta2=r2=0.266, and when average strength of the alternative hypothesis was included as a covariate $F(2, 22)=3.59$, $p=0.045$, partial eta2= 0.246 (value for one user was lost). Variances were not equal across groups, but *t*

Table 5 Inclusion of positive, negative, and neutral information by condition.

	Heterogeneous	Homogenous	Solo	All conditions
% All items	52.9	42.1	54.7	49.9
% Positive, relevant	74.07	59.6	73.74	69.14
% Negative, relevant	65.97	59.03	80.56	68.52
% Neutral, irrelevant	23.66	14.7	21.15	23.66

Table 6 Coverage of evidence by triad.

	Heterogeneous (%)	Homogenous (%)
Interacting triad (actual interacting groups)	97.30	83.70
Statisticized triad (post-hoc combinations)	97.90	97.30

We measure if any group member included the data in their matrix. First row: average coverage of interacting groups. Second row: average coverage of two statisticized groups, post-hoc combinations of three individuals' data, where each triad includes individuals homogeneously or heterogeneously biased.
Coverage measure: % of 64 pieces of relevant evidence included

tests comparing the Homogeneous and Solo Conditions, with and without assuming equal variance, also found a significant difference between conditions, at $p=0.025$, ts(15)=2.5 to 2.6. (A related MANOVA with condition and initial bias as factors also showed significant effects of condition, but not of initial Preferred Hypothesis or their interaction.)

6.4. Amount, source and type evidence

Table 5 shows the amounts and types of evidence users included in their matrix. First, users include very similar proportions of negative and positive evidence, suggesting that in these conditions users see the relevance of disconfirmatory evidence (independent of what hypothesis it disconfirms). Second, users discriminate strongly between relevant and irrelevant evidence. Third, there are no condition differences, though users in the Homogeneous Condition show nonsignificantly lower rates of inclusion. A repeated measures analysis (and MANOVA) on percent of items included in the matrix with item type (positive, negative, and neutral) and condition as factors showed a significant effect of item type, $F(2, 48)=120.9$, $p<0.001$, but not of condition, $F(2, 24)=1.28$, $p=0.295$, eta2=0.097), nor their interaction, $F(2, 48)=1.33$, $p=0.272$, observed eta2=0.137.

One possible advantage of working in an interacting group is that interacting individuals might collectively view more evidence than individuals working alone. For example, this might come about if the group split up what to look at, or if individuals were inclined to look for different types of information. We looked descriptively at coverage (the amount of relevant evidence included in at least one group member's matrix) by actual interacting individuals versus by sets of noninteracting individuals. The noninteracting sets were triads of individuals grouped together post-hoc, where each triad included homogeneously or heterogeneously biased individuals (see statisticized groups in Schultz-Hart et al. 2000). Because we had only a few groups, we simply note the triad coverage. Coverage of relevant evidence is very high overall at the triad level (94%), though lowest in the homogeneous, interacting group (Table 6).

6.5. CACHE-log: online tracking

Online logging can show patterns in viewing evidence. We successfully logged three of the Homogeneous groups, and two of the Heterogenous groups, showing each occurrence of opening a piece of evidence (i.e., to read it), and each time a piece of evidence was added to a matrix. The nine participants in the Homogeneous condition read 34.7 items on average, compared to 45.5 items averaged by the six participants in the Heterogeneous Condition. Homogeneous participants averaged 39.2 items added, compared to the Heterogeneous participants, who averaged 53.2. This suggests that these participants added more evidence than they read, consistent with a 'breadth first' strategy, of including unread material in the matrix; we had not anticipated participants developing this strategy.

Online logs provide data about what evidence is accessed, and evidence which could only be accessed through a partner is of particular interest. During Blocks 2 and 3, certain pieces of evidence could only be accessed by a given analyst through viewing their partners' matrices or through their ticker reports on partner activity. Groups and individuals varied enormously in the number of times they borrowed a piece of evidence from a partner (Borrow event). The two heterogeneous groups had 3 and 53 Borrow events, with Borrowing distributed across participants and evidence. The Homogeneous Groups had 2, 5, and 16 Borrow events respectively, with the borrowing in the high-use group performed primarily by one participant, who was rereading a small set of evidence.

6.6. Usability ratings and open comments

Participants evaluated the tools and the task through five-point usability and usefulness rating scales, and through open-ended questions about the tools and task. Overall, ratings were positive (mean=4.04) on the 26 items focusing on the CACHE-A system, grouped into six composite measures. On all six measures, the confidence interval was above 3, the scale midpoint. Comments were favorable and participants were engaged by the task and tools. Responses were generally similar across conditions; points of condition and individual difference are mentioned here.

With respect to the task, participants in the Solo condition judged that there was sufficient time for the task, while participants in the interactive conditions (Homogeneous and Heterogeneous) judged that they had too little time, presumably because of the additional process costs of managing group interaction. This was supported by spontaneous mention of time-pressure in the open-ended questions by 8 of 18 interacting participants. None of the nine solo participants mentioned time pressure in the open-ended questions.

The Ticker window handled task management and instructions from the experimenter (advancing through the task) for all users but included partner communication only for the two interacting conditions. Conditions differed, due to very low ratings by Solo (mean 2.3) and Homogeneous (mean 2.2) conditions, but

high ratings by the Heterogeneous condition (mean 3.8). Heterogeneous analysts tolerated the interruptions from the ticker for the value of timely partner notification, even though Homogeneous and Heterogeneous analysts rated partner input similarly on other measures. Fourteen of 27 participants wanted more control of how information was displayed in the matrix, both for individual purposes and to better compare content with partner matrices. Ten participants asked for better window management or fewer windows. These and other suggestions can be used to inform the development of CACHE and similar systems, and for the refinement of our experimental task for studying computer-supported collaborative analysis.

Groups varied in how much they used Chat as well as how much they used partner matrices (for getting evidence). The group that chatted least pulled evidence from partner matrices the most. Individuals differed in how they used their matrix. We assumed that users would read evidence, evaluate its relevance, and if relevant include in their matrix. Several users, however, included almost everything, and then dropped or gave low relevance to unimportant information. This may indeed have been a very efficient method for systematically reviewing a large amount of evidence.

7. Discussion

7.1. Summary of results and significance

We studied how group composition affects the performance of group members in a computer-supported intelligence analysis task. After inducing an initial belief, we assessed strength of confirmation bias, that is, adherence to that belief more strongly than warranted by evidence. We found significant group differences. Individuals in a Heterogeneous group and those who worked alone debiased strongly and similarly as they were exposed to corrective evidence; their belief ratings in the preferred hypothesis dropped 17%, from about 60% to 43% and no longer differed significantly from the norm of 33%. In contrast, individuals in a Homogeneous group showed increasing bias of 15%, from 57% to 72%; rather than responding to debiasing evidence available over blocks, the exposure to evidence and judgments of similarly biased others strengthened the initial belief.

Similar differences in confirmation bias due to group condition were found in two additional measures: (1) the importance score (Table 4 row b), measured as the difference between importance assigned to positive evidence items supporting the preferred hypothesis versus those assigned to the alternatives; and (2) the system-computed strength score (Table 4 row d), which provides a weighted sum of evidence-strength for the preferred hypothesis. Note that both these measures incorporate any bias in the inclusion and assessment of individual pieces of evidence, but not in their integration, as the global Belief score does (or judgment at the level of the hypotheses).

To summarize, we found two phenomena concerning the effect of working in a group versus working alone in this computer-supported environment: (1) confirmation bias was no worse for an individual when working in a heterogeneous group than when working alone (no net process cost here) but (2) confirmation bias was much worse when working in a homogeneous group than working alone. Schulz-Hardt and collaborators had shown with face-to-face groups that confirmation bias is strengthened in homogeneous groups and tempered in heterogeneous groups. Our findings replicate and extend theirs, to computer-supported distributed work, a more complex judgment task, and effects of a group on individual members. Additionally, we collected a broader range of measures of bias (see below).

We make two observations about the context in which we found the significant, parallel effects on belief, strength, and importance scores, among the multiple measures we assessed. First, the overall *occurrence* of confirmation bias was widespread and significant across multiple measures, as summarized in Table 4 (see averages with "*"). Second, the pattern of differences due to group condition was similar across all the measures, whether or not effects were significant (see averages with "**"). Because sample size was small and measures differed in variability, it is difficult to tell whether measures differed in significance because of these sensitivity factors or because patterns in some but not other measures reflect true population differences.

Additional research with larger samples would allow assessing the robustness of these effects. In particular, it may be valuable to use this fine-grained data collection and analysis method to "deconstruct" the biased group decision process into stages or sub-processes – for example, item selection, assessment of importance, assessment of item–hypotheses consistency relation, and integration across items and hypotheses. To the extent that the different measures differentially reflect identifiable and distinct sub-processes, we can measure what portion of the final confirmation bias emerges in each component process, and how group composition affects individuals' behavior at each stage. Ultimately, useful implications for the design of debiasing tools can be derived on the basis of the findings.

Our results extend those by Schultz-Hart et al. (2000) in three ways. Most importantly, we investigate the effects of group composition when participants' work environment provides computational support for both individual and collaborative aspects of the task. Second, we study the performance of groups working on a substantially more complex problem: a larger body of evidence (80 positive, negative, or neutral statements, rather than ten positive or negative statements in Schultz-Hardt et al.) and three relevant hypotheses (rather than two as in Schultz-Hart et al. 2000). The number of hypotheses qualitatively changes the problem structure: with only two, A-or-not-A hypotheses, any evidence against one hypothesis is support for the other, while with three the relations among hypotheses are more complex. Complexity is critical for learning about realistic collaboration problems. In addition, the need for computer support is

greater for more complex problems. For example, the benefits of explicit representation and updating of multiple alternative hypotheses (as CACHE-A provides) may be particularly valuable when a problem concerns more than two alternatives. Third, we extend the scope of the task and provide additional measures of bias throughout the task: evidence items inclusion, assessment, and integration in a final summative belief (expressed at the level of the hypotheses). Thus, we measure not only the bias in evidence item selection (as in Schultz-Hart et al. 2000, and in our Table 4 line a), but also the bias in evidence item assessment (see Table 4 lines b–d) and, most importantly, in the overall belief about the three hypotheses (see Table 4 line e). In addition, we extend the task across multiple blocks and track bias after each block rather than only once.

7.2. Hypotheses about computer support for collaboration: CACHE-A's role

Prior research has suggested that collaboration can reduce bias, but at the same time it brings significant process costs. The challenge for CSCW research is to find ways to preserve the benefits and minimize the process costs. We did not assess the effects of the CACHE-A system on collaboration, because we did not compare collaboration with and without support. However, our findings lead to a more complicated hypothesis about the role for collaborative technology. The results suggest that CACHE-A did reduce the costs of communicating and sharing information among group members, but that reduction of these costs played out differently in differently composed groups. We hypothesize that CACHE-A could reduce collaborative costs. Support for this hypothesis comes from the similarity between performance of the individuals in Heterogeneous Groups and of individuals who worked alone. Given the pervasive occurrence of process costs in face to face groups, this lack of cost for Heterogeneous groups working with CACHE-A suggests that CACHE-A facilitated interaction. This lack of damage to Heterogeneous Group performance held despite the fact that collaborating individuals had more to learn, more windows to manage, and more information bidding for their attention than individuals working alone for the same time, and despite the fact that collaborating individuals perceived their workload as higher than individuals working alone.

We hypothesize several ways in which CACHE-A benefited performance and reduced collaborative costs. We suspect that learning the ACH method and provision of an efficient tool for building ACH matrices helped reduce bias at an individual level. We suspect that the transparent, shared form of the CACHE-A matrices made them easy to understand when shared, and the interface made it low cost to access the shared matrix. The ticker and chat provided complementary low cost means for tracing progress, sharing information, and discussing interpretations. This support facilitated communication. We hypothesize, however, that the effects on debiasing from aiding group communication play out differently depending on the composition of the computer-supported group: when

groups are diverse CACHE-A accentuates the beneficial debasing effect of group composition on performance, compensating for time spent managing communication; in contrast, when group members share the same bias, CACHE-A may actually accentuate the group effects that magnify bias. This possibility merits direct investigation.

Two general classes of factors affect group decision-making: the properties of the group and the properties of the medium (see Section 2.4). Our experiment manipulated the group properties but not the medium. Thus we can only speculate about the effects that the CACHE-A medium may have on group decision making, drawing on prior studies that have compared face-to-face and computer-mediated groups. Kiesler and Sproull (1992) reviewed prior research and noted that removing facial and verbal cues in computer-mediated communication and decision-making introduces the benefits of a more open and blunt discussion style and less inequality of participation in discussions, but also reduces awareness, slows decision processes, and increases the effort to reach consensus because the members use more explicit verbal statements of their positions. Therefore, the reduced amount of social cues among CACHE-based collaborators may remove sources of bias such as status cues (e.g., Kiesler and Sproull 1992), keep collaborators more focused on the content to be shared for more efficient information sharing (e.g., Convertino et al. 2008), but require greater effort for maintaining awareness and reaching final judgment. This tradeoff may be preferable for domains where the accuracy and completeness of the decision are particularly critical, such as intelligence analysis.

7.3. Collaborative analysis and requirements for its support

With others, we are struck by the different forms and degrees of collaboration involved in intelligence analysis and by the need for fluent transitions among these forms of collaboration. Our task required participants to iteratively change between individual and collaborative work modes over time. Further, within this structure, different groups took different approaches. Some interacted only through viewing partners' matrices when they were available, with little or no discussion or questioning. Others carried out discussions, tried to divide work, or sought group-level agreement on the final ratings of hypotheses.

A key aspect of analysts' needs may be for flexibility in working in a variety of collaborative modes and in transitioning fluently among these. Variation in collaboration mode can be represented as a multidimensional space varying in the types and strengths of interaction. A specific activity forms a trajectory over time through this space, as mode of work changes across a task. A multidimensional and continuous framework, rather than a categorical one (such as synchronous versus asynchronous), affords a higher-resolution picture of varying work activity. For example, work activity varies in how tightly coupled participants are in time and in intention. Interactions with tight *temporal coupling* involve

short delays, as in a conversation as opposed to delayed communications as in letters or publications. In interactions with tight *intent coupling* participant interaction is guided by mutually known, shared goals as opposed to incidental, indirect benefits as in shared tagging. Even when "working alone" an individual's activity is embedded in a larger group context. A continuous view of activity space may facilitate design of systems which can be used over a dynamic variety of work modes. For example, tools such as CACHE-A's shared matrices facilitate sharing representations constructed as part of individual work and this capability may be particularly helpful for effective collaboration.

7.4. Design for collaborative support

Design of CSCW systems requires a balance between accommodating current practices and envisioning how technology can enable new and better ways to work. Intelligence analyst are "extreme" information workers under multiple pressures, including the need for varying and distributed modes of work, flexibility, processing massive amounts of information, meeting high performance expectations, and quick turnaround. Designing for the needs of analysts may thus address a particularly forward-looking case, of increasingly wide applicability to other information workers in the future.

Seminal research work on bias from the 1970s and 1980s changed how individual decision-making under uncertainty and risk was later understood and supported (see impact of Tversky and Kahneman's work on Decision-Making and Economics research). Future research on group bias in computer-supported conditions may hold a similar opportunity. New technologies such as search engines, Wikis, recommender systems, tools for democracy, and tools for scientific collaboratories are introducing new forms of collective intelligence and impacting people's decisions in new ways. The CACHE-A software used in this study is a more explicit case, part of the same trend, supporting collaborative analysis. Findings about the component processes of group bias can guide the design and evaluation of new collaborative technology that effectively enhances the quality of group decisions.

The interplay between experimental findings and design ideas can be illustrated with our work. Considering our findings from the perspective of a continuous, multidimensional space suggests a novel function for collaborative software. Rather than assuming a fixed mode of interaction, "collaboration agents," such as those implemented in simple form in CACHE-A, could monitor individuals' activity and advise where and when varied forms of collaboration might be useful. Software such as CACHE-A, which has information about participants' beliefs, might be used to advise both on the composition of groups, and to modulate communication depending on the heterogeneity or diversity of opinion present in the group. For example, it might recommend forming an interacting group with heterogeneous members, or separating a group with homogeneous

beliefs to optimize productive collaboration. Shrager (in review) describes the more general CACHE framework and the way that the full CACHE vision builds upon this fluidity to enable communities of investigators (scientists or analysis) to collaborate in collective multistage inference.

We believe that the availability of tools such as CACHE will enable analysts to accomplish their current tasks and activities more efficiently, and also will change the nature of the work that analysts can perform. Further lab experiments and fieldwork focusing on the effects of such technology will be important in order to understand both their efficacy and how social and individual processes play out within the context of a particular computer-support system.

Acknowledgements

This research was funded by the Novel Intelligence from Massive Data program, under contract no. MDA904-03-C-0404 and by the Office of Naval Research no. N00014-96-C-0097. We thank Stuart Card and the UIR researchers at PARC for their feedback, and MITRE Corporation, Naval Postgraduate School, PARC employees, and Stanford students for contributing to this study.

1. Appendix

Additional screen shots show details of the CACHE-A interface. Examples show content from training phase.

Figure 7 ACH Matrix row: tabular interface to specify weight or importance of one evidence item and its consistency relation to the three hypotheses.

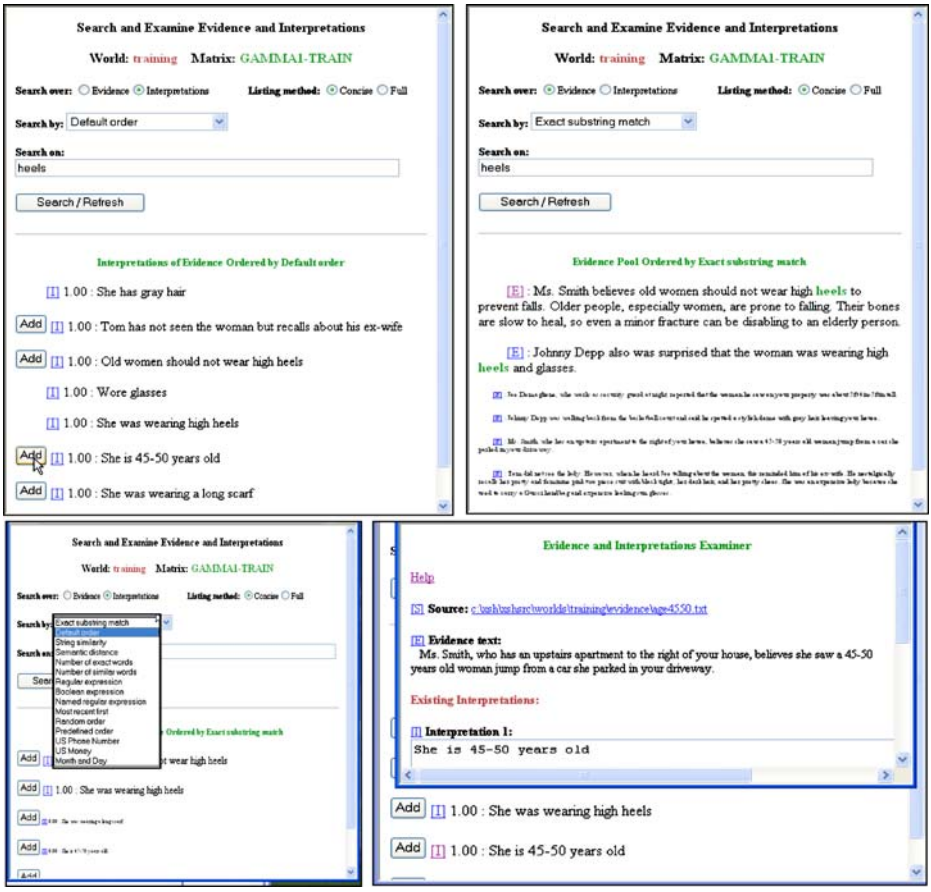

Figure 8 Search tool. The two top figures show the search tool. In the tool, the analyst specifies **a** if to search over interpretations (*top left*) or items content (*top right*), **b** the search keywords, **c** the listing method: i.e., concise vs. full results, and **d** the keyword-result marching method (*bottom left*). A specific item can be viewed by clicking on the "[I]" (*top left*) link or "[E]" link (*top right*) at the bottom of the search tool. An evidence viewer (*bottom right*) will open for this item as a result.

Figure 9 Evidence viewer. The window displays the content of an evidence item and its interpretations. Each item has one default interpretation. The analyst can edit the current interpretation or add a new interpretation using the text field under the evidence text.

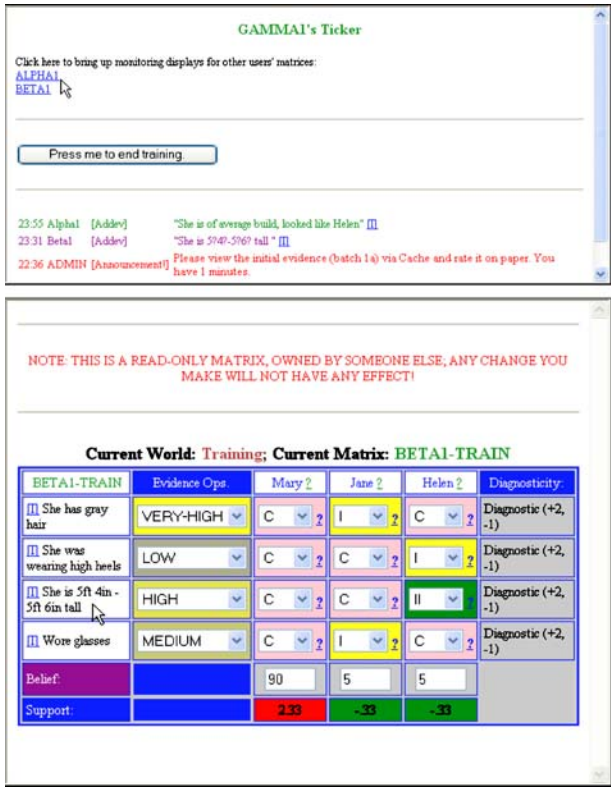

Figure 10 Ticker and partner's matrix. The analyst can open read-only views of collaborators' ACH matrices by clicking on the top links in the Ticker, while CACHE-A is in collaborative mode. The links "ALPHA1" and "BETA1" in the ticker, the top figure, give GAMMA1 access to her/his partner matrices. The bottom figure shows the read-only view of BETA1's matrix that would open on her/his screen.

References

Adelman, L., Tolcott, M.A. and T.A. Bresnick (1993): Examining the effect of information order on expert judgment, *Organizational Behavior and Human Decision Processes*, vol. 56, pp. 348–369.

Arnott, D. (2006): Cognitive Biases and Decision Support Systems Development: A Design Science Approach. *Information Systems Journal*, vol. 16(1), pp. 55–78 doi:10.1111/j.1365-2575.2006.00208.x.

Benbasat, I. and J. Lim (2000): Information Technology Support For Debiasing Group Judgments: An Empirical Evaluation. *Organizational Behavior and Human Decision Processes*, vol. 83, pp. 167–183 doi:10.1006/obhd.2000.2905.

Benjamin, K. (1990): Why We Still Use Our Heads Instead of Formulas: Toward an Integrative Approach. *Psychological Bulletin*, vol. 107(3), pp. 296–310 doi:10.1037/0033-2909.107.3.296.

Bornstein, B.H. and A.C. Emler (2001): Rationality in Medical Decision Making: A Review of the Literature On Doctors' Decision-making Biases. *Journal of Evaluation in Clinical Practice*, vol. 7(2), pp. 97–107 doi:10.1046/j.1365-2753.2001.00284.x.

Camerer, C.F. and E.J. Johnson (1991): The Process–Performance Paradox in Expert Judgment: How Can Experts Know so Much and Predict so Badly? In K.A. Ericsson and J. Smith (eds): *Towards a General Theory of Expertise: Prospects and Limits*New York: Cambridge University Press, pp. 195–217.

Card, S.K. (2005): The Science of Analytical Reasoning. In J.J. Thomas and K.A. Cook (eds): *Illuminating the Path: the Research and Development Agenda for Visual Analytics*IEEE CS: Los Alamitos, CA.

Chapman, G.B., G.R. Bergus and A.S. Elstein (1996): Order of Information Affects Clinical Judgment. *Journal of Behavioral Decision Making*, vol. 9(3), pp. 201–211 doi:10.1002/(SICI)1099-0771(199609)9:3<201::AID-BDM229>3.0.CO;2-J.

Cheikes, B.A., M.J. Brown, P.E. Lehner and L. Alderman (2004): *Confirmation Bias in Complex Analyses*. Technical Report No. MTR 04B0000017. Bedford, MA: MITRE.

Cho, H.-K. (2004): *The effect of Delphi Structure on Small and Medium-sized Asynchronous Groups*. Ph. D. Dissertation Thesis, New Jersey Institute of Technology, Information Systems Department.

Cho, H.-K. and M. Turoff (2001): Debiasing Group Judgments through Computerized Delphi Systems. In *Proceedings of AMCIS 2001*, August, Boston, MA.

Convertino, G., H.M. Mentis, P. Bhambare, C. Ferro, J.M. Carroll and M.B. Rosson (2008): Comparing Media in Emergency Planning. In *Proceedings of the 5th International ISCRAM. Conference*. Washington, DC, USA, May 4–7, 2008.

Cook, M.B. and H.S. Smallman (2007): Collaborative intelligence analysis: Debiasing through graphical evidence layout. In: *Proceedings of the 51st Annual Meeting of the Human Factors and Ergonomics Society*, Baltimore, MD, pp. 16–20.

Cummings, J.N. (2004): Work Groups, Structural Diversity, and Knowledge Sharing in a Global Organization. *Management Science*, vol. 50(3), pp. 352–364.

Davis, E.B. and R.H. Ashton (2002): Threshold Adjustment in Response to Asymmetric Loss Functions: The Case of Auditors' "Substantial Doubt" Thresholds. *Organizational Behavior and Human Decision Processes*, vol. 89, pp. 1082–1099 doi:10.1016/S0749-5978(02)00009-2.

Dennis, A.R. and J.S. Valacich (1993): Computer Brainstorms: More Heads are Better Than One. *The Journal of Applied Psychology*, vol. 78(4), pp. 531–537 doi:10.1037/0021-9010.78.4.531.

Dennis, A.R., K.M. Hilmer and N.J. Taylor (1997): Information Exchange and Use in GSS and Verbal Group Decision Making: Effects of Minority Influence. *Journal of Management Information Systems Archive*, vol. 14(3), pp. 61–88.

Fjermestad, J. and S.R. Hiltz (2001): An Assessment of Group Support Systems Research: Methodology. *Journal of Management Information Systems*, vol. 15(3), pp. 7–149.

Fugelsang, J.A., C.B. Stein, A.E. Green and K.N. Dunbar (2004): Theory and Data Interactions of the Scientific Mind: Evidence From the Molecular and the Cognitive Laboratory. *Canadian Journal of Experimental Psychology*, vol. 58(2), pp. 86–95 doi:10.1037/h0085799.

Gallupe, B.R., L.M. Bastianutti and W.H. Cooper (1991): Unblocking Brainstorms. *The Journal of Applied Psychology*, vol. 76, pp. 137–142 doi:10.1037/0021-9010.76.1.137.

George, J.F., K. Duffy and M. Ahuja (2000): Countering the Anchoring and Adjustment Bias with Decision Support Systems. *Decision Support Systems*, vol. 29(2), pp. 195–206 doi:10.1016/S0167-9236(00)00074-9.

Gettys, C.F., C. Kelly III and C.R. Peterson (1982): The Best-guess Hypothesis in Multistage Inference. In D. Kahneman, P. Slovic and A. Tversky (eds): *Judgment Under Undertainty: Heuristics and Biases*New York: Cambridge University Press, pp. 370–377.

Gigone, D. and R. Hastie (1996): The Impact of Information on Small Group Choice. *Journal of Personality and Social Psychology*, vol. 72, pp. 132–140 doi:10.1037/0022-3514.72.1.132.

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

Hightower, R. and L. Sayeed (1995): The Impact of Computer-mediated Communication Systems on Biased Group Discussion. *Computers in Human Behavior*, vol. 11(1), pp. 33–44 doi:10.1016/0747-5632(94)00019-E.

Hogarth, R.M. and Hillel J. Einhorn (1992): Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, vol. 24(1), pp. 1–55 doi:10.1016/0010-0285(92)90002-J.

Johnston, R. (2003): Reducing Analytic Error: Integrating Methodologists into Teams of Substantive Experts. *Studies in Intelligence*, vol. 47(1), pp. 57–65.

Johnston, R. (2005): *Analytic Culture in the U.S. Intelligence Community: An Ethnographic Study.* Washingon, DC: Central Intelligence Agency, Center for the Study of Intelligence.

Johnson, E.M. and S.M. Halpin (1974): *Multistage Inference Models For Intelligence Analysis. Report AD 785-639*. Arlington, VA: Army Research Institute for the Behavioral and Social Sciences(Distributed by US Department of Commerce, Springfield, VA).

Kahneman, D., P. Slovic and A. Tversky (eds) (1982): *Judgment under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.

Keefer, Donald L., Craig W. Kirkwood and James L. Corner (2004): Perspective on Decision Analysis Applications, 1990–2001. *Decision Analysis*, vol. 1(1), pp. 4–22 doi:10.1287/deca.1030.0004.

Kerr, N.L. and S.R. Tindale (2004): R.S. Group Performance and Decision Making. *Annual Review of Psychology*, vol. 55, pp. 623–655 doi:10.1146/annurev.psych.55.090902.142009.

Kerr, Norbert L., R.J. MacCoun and G.P. Kramer (1996): Bias in Judgment: Comparing Individuals and Groups. *Psychological Review*, vol. 103(4), pp. 687–719 doi:10.1037/0033-295X.103.4.687.

Kerstholt, J.H. and J.L. Jackson (1998): Judicial Decision Making: Order of Evidence Presentation and Availability of Background Information. *Applied Cognitive Psychology*, vol. 12, pp. 445–454.

Kiesler, S. and L. Sproull (1992): Group Decision Making and Communication Technology. *Organizational Behavior and Human Decision Processes*, vol. 52, pp. 96–123 doi:10.1016/0749-5978(92)90047-B.

Klayman, J. and Y.-W. Ha (1987): Confirmation, Disconfirmation, and Information in Hypothesis Testing. *Psychological Review*, vol. 94(2), pp. 211–228 doi:10.1037/0033-295X.94.2.211.

Kraut, R.E. (2003): Applying Social Psychological Theory to the Problems of Group Work. In John M. Carroll (ed): *HCI Models, Theories, and Frameworks: Toward a Multidisciplinary Science*New York: Morgan Kaufmann, pp. 325–356.

Larson, J.R. Jr., P.G. Foster-Fishman and C.B. Keys (1994): Discussion of Shared and Unshared Information in Decision-making Groups. *Journal of Personality and Social Psychology*, vol. 67, pp. 446–461 doi:10.1037/0022-3514.67.3.446.

Lim, L.-H. and I. Benbasat (1997): The Debiasing Role of Group Support Systems: An Experimental Investigation of the Representativeness Bias. *Int. J. Human–Computer Studies*, vol. 47, pp. 453–471 doi:10.1006/ijhc.1997.0137.

Lipshitz, R., G. Klein, J. Orasanu and E. Salas (2001): Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making*, vol. 14(5), pp. 331–352 doi:10.1002/bdm.381.

Neale, D.C., J.M. Carroll and R.M. Beth (2004): Evaluating Computer-supported Cooperative WOrk: Models And Frameworks. In *Proceedings of the ACM conference on Computer Supported Cooperative Work*, Chicago, IL.

Nunamaker, J.F. Jr., A.R. Dennis, J.S. Valacich, D.R. Vogel and J.F. George (1991): Electronic Meeting Systems to Support Group Work. *Communications of the ACM*, vol. 34(7), pp. 40–61.

Perrin, B.M., B.J. Barnett, L. Walrath and J.D. Grossman (2001): Information Order and Outcome Framing: An Assessment of Judgment Bias, In a Naturalistic Decision-Making Context. *Human Factors*, vol. 43(2), pp. 227–238 doi:10.1518/001872001775900968.

Pirolli, P., T. Lee and Stuart K. Card (2004): *Leverage Points for Analyst Technology Identified through Cognitive Task Analysis Technical Report*. Palo Alto, CA: PARC.

Pirolli, P., L. Good, J. Heiser, J. Shrager and S. Hutchins (2005): *UIR Technical Report*. Palo Alto, CA: PARC.

Reagan-Cirincione, P. (1994): Improving the accuracy of group judgment: a process intervention combining group facilitation, social judgment analysis, and information technology. *Organizational Behavior and Human Decision Processes*, vol. 58, pp. 246–270 doi:10.1006/obhd.1994.1036.

Russell, D.M., M.J. Stefik, P. Pirolli, and S.K. Card (1993): The Cost Structure of Sensemaking. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, April 24–29, 1993, Amsterdam, The Netherlands, pp. 269–276.

Scholtz, J., E. Morse and P.S. Michelle (2006): Evaluation metrics and methodologies for user-centered evaluation of intelligent systems. *Interacting with Computers*, vol. 18, pp. 1186–1214 doi:10.1016/j.intcom.2006.08.014.

Schultz-Hart, S., D. Frey, C. Lüthgens and S. Moscovici (2000): Biased Information Search in Group Decision Making. *Journal of Personality and Social Psychology*, vol. 78(4), pp. 655–669 doi:10.1037/0022-3514.78.4.655.

Shanteau, J. (1992): Competence in Experts: The Role of Task Characteristics. *Organizational Behavior and Human Decision Processes*, vol. 53, pp. 252–266 doi:10.1016/0749-5978(92)90064-E.

Shrager, J. (2005): CACHE; The Collaborative Analysis of Competing Hypotheses Environment [computer software]. Palo Alto, CA: Xerox Palo Alto Research Center.

Shrager, J., D. Billman, G. Convertino, J.P. Massar and P. Pirolli (in review): CACHE: Web-based support for distributed multi-stage inference.

Smallman, H.S. (2008): JIGSAW – Joint Intelligence Graphical Situation Awareness Web for Collaborative Intelligence Analysis. In M.P. Letsky, N.W. Warner, S.M. Fiore and C.A.P. Smith (eds.): *Macrocognition in Teams: Theories and Methodologies*. Ashgate Publishing Limited, Hampshire, England, pp. 321–337 (in press).

Stasser, G. and W. Titus (2003): Hidden Profiles: A Brief History. *Psychological Inquiry*, vol. 14(3–4), pp. 302–311.

Straus, S.G. and J.E. McGrath (1994): Does the Medium Matter? The Interaction of Task Type and Technology on Group Performance and Member Reactions. *The Journal of Applied Psychology*, vol. 79(1), pp. 87–89 doi:10.1037/0021-9010.79.1.87.

Tolcott, M.A., F.F. Marvin and P.E. Lehner (1989): Expert Decisionmaking in Evolving Situations. *IEEE Transact. on Systems. Man, and Cybernetics*, vol. 19(3), pp. 606–615 doi:10.1109/21.31066.

Trent, S.A., E.S. Patterson and D.D. Woods (2007): Challenges for Cognition. *J. of Cognitive Engineering and Decision Making*, vol. 1(1), pp. 75–97.

Tversky, A. and D. Kahneman (1974): Judgments under Uncertainty. Heuristics and Biases. *Science*, vol. 185, pp. 1124–1131 doi:10.1126/science.185.4157.1124.

van Knippenberg, D. and M.C. Schippers (2007): Work Group Diversity. *Annual Review of Psychology*, vol. 58, pp. 515–541 doi:10.1146/annurev.psych.58.110405.085546.

Wason, P.C. (1960): On the Failure to Eliminate Hypotheses in a Conceptual Task. *The Quarterly Journal of Experimental Psychology*, vol. 12, pp. 129–140 doi:10.1080/17470216008416717.