Informing the development of a fraud prevention toolset through a situated analysis of fraud investigation expertise

Final author draft only. The final copyedited article may differ slightly from this manuscript version.

Thomas C. Ormerod

Department of Psychology, Lancaster University

Lancaster, LA1 4YF, UK

Tel: +44 (0)1524 593164; Fax: +44 (0)1524 593744
Email: T.Ormerod@lancaster.ac.uk

Linden J. Ball*

Department of Psychology, Lancaster University

Lancaster, LA1 4YF, UK

Tel: +44 (0)1524 593470; Fax: +44 (0)1524 593744
Email: L.Ball@lancaster.ac.uk

Nicola J. Morley

Department of Psychology, Lancaster University

Lancaster, LA1 4YF, UK

Tel: +44 (0)1524 593698; Fax: +44 (0)1524 593744
Email: Nicki.Morley@btinternet.com

---

*Corresponding Author. Email: L.Ball@lancaster.ac.uk

Abstract

Insurance fraud is a growing problem. This paper describes a project that aimed to specify and develop a suite of computer-based tools to support the early detection and subsequent investigation of potentially fraudulent claims. System requirements were informed by ethnographic studies focusing on: (1) understanding current fraud detection practices; and (2) characterising fraud investigation expertise. Tools were designed that sift claims for potential problem cases and assist in the processes of investigation and detection of new fraud types by providing anomaly capture, argumentation and visualisation environments. The resulting tools capitalise upon expertise and embody processes that can subsequently be used by inexperienced claims handlers to detect and deal with fraudulent claims.

Keywords

Insurance fraud; cognitive ethnography; anomaly capture; computer-based tools; decision support systems

1. Introduction

Insurance fraud presents a major and growing problem that is felt most immediately by the insurance industry. For example, the Association of British Insurers (2006) has estimated that the value of fraudulent claims presented to insurers in the UK is approximately 1.5 billion pounds per annum. This figure suggests that internationally the cost of such fraud is likely to be immense. Insurance fraud not only impacts upon the insurance industry but also affects all policyholders through increased premiums and, in some cases, increased exclusions and difficulties in obtaining insurance cover. Insurance fraud is also a complex and multi-faceted problem. It varies in scale from inflated claims for genuine incidents through to systematic multi-person 'scams' that involve staged accidents, thefts, and so on. In addition, insurance fraud is dynamic: as one scam is uncovered, a new one takes its place. Furthermore, the need to maintain customer loyalty and efficient sales practices means that the kinds of information checks that can be carried out at policy inception and during the handling of a claim are limited and time-pressured.

Against this backdrop the insurance industry is actively seeking measures that can speed-up the processing of legitimate claims whilst also increasing the number of fraudulent claims that are detected. Indeed, over the past ten years or so a number of technological interventions have been developed and put into place. Yet, to date, there is no single satisfactory technological approach available for dealing with fraud. In this paper, we overview the key empirical and technological aspects of FRISC (Fraudulent Insurance Claims), a programme of research that we have been involved with that aims to provide computer-based solutions to reduce fraud and assist in the detection of new fraud types.
Two main approaches to tackling fraud through technology can be identified. The first, and perhaps most common, is the use of data-mining methods whereby systems trawl databases of client and claim information, cross-checking with external data sources to identify

inconsistencies and anomalies associated with claims (see Yue, Wu, Wang, Li, & Chu, 2007). Perhaps the best known data-mining system in use in the UK banking and insurance sector is the 'Hunter II' fraud-detection system developed by Experian Decision Analytics (2008). Applications of data mining techniques have also been extended to fraud-detection in a wide variety of financial domains, including credit card fraud (e.g., Chan, Fan, Prodromidis, & Stolfo, 1999), cellular cloning fraud (e.g., Fawcett & Provost, 1997), and telephone calling fraud (e.g., Cox, Eick, Willis, & Brachman, 1997), attesting to the generality of this approach.

Arguably the main strength of data mining relates to the ability to process large volumes of claim-relevant information across multiple sources to reveal potential frauds. However, because data mining searches information that is provided by customers and entered by human operators, the risk of false positives is high. Mismatches between name, title, zip or postal code, policy number, inception date, and so on, are anomalies, yet many arise through innocent data entry slips, memory failures and so on. Moreover, the anomalies that are identified require expert interpretation: two different claimant names at the same address may or may not be more indicative of a fraudulent claim than the same name at two different addresses (see Morley, Ball, & Ormerod, 2006).

The second technology-based approach to tackling fraud uses methods aimed at 'profiling' likely fraudsters, including techniques such as voice stress analysis (e.g., Hollien, Geison, & Hicks, 1987; Horvath, 1982). The value of these kinds of profiling approaches reside both in their ability to detect repeat fraudsters and also to dissuade opportunists. However, techniques like voice stress analysis require direct contact with the insured, and contact slows the process of claims handling. Moreover, a large proportion of insurance business is conducted through third party brokers and increasingly across the Internet.

An alternative technological approach to tackling fraud is to provide tools that support an expert's existing fraud detection knowledge and practices. Some generic visualisation tools are used to this end such as the i2 system (see i2 Solutions, 2008), but the value of such tools is restricted to the identification of large and organised frauds, and their use requires intensive training. We believe that insufficient attention has been given to understanding the nature of fraud detection and investigation expertise itself. This is perhaps surprising, given that many of the larger insurance companies have specialist fraud detection and investigation units that typically show a large return on investment. Moreover, there is evidence from studies of crime investigation across a range of domains (including scene of crime analysis, counter-terrorism and hostage negotitation) that specialist investigators offer the kinds of expertise that can be capitalised upon in the development of investigative processes and support tools (Ormerod, Barratt, & Taylor, 2008).

The main aim of our research is to study existing fraud detection and investigation expertise in depth and to use the findings of these studies to specify and prototype a range of fraud detection tools. Our research adopts an ethnographic approach to empirical study (see Hammersley & Atkinson, 1995) using techniques such as interviewing, observation and practical hands-on experience to characterise the process of fraud detection and investigation. Ethnography has been increasingly applied to systems design (e.g., Anderson, 1994; Ball & Ormerod, 2000a, 2000b; Button, 2000; Crabtree, Nichols, O'Brien, Rouncefield, & Twidale, 2000; Hughes, Sommerville, & Bentley, 1993). It has a number of advantages over other system specification approaches. For example, it is non-invasive and captures the complexities of social interaction around technological systems in real working environments. In addition, the long-term, situated and intensive nature of this approach lends itself to understanding the complex environment in which insurance claims are detected and investigated. As well as characterising fraud expertise to inform the design of our decision support tool, our research focused on understanding the environment in which these tools would be utilised.

The remainder of the paper is structured as follows. First, we provide an overview of the ethnographic studies that we undertook in major insurance companies and loss adjustment firms. In describing this ethnographic research we focus on key observations that informed our understanding of expert fraud detection and investigation practices and which clarified how best to support such expertise through technological tools. We then discuss the software prototypes whose design was informed by the ethnographic studies. One prototype acts as a 'sieve' to identify claims that carry a high fraud risk. The other uses visualisation, argumentation and anomaly evaluation capabilities to support the processes of investigating and repudiating potentially fraudulent claims.

2. Ethnographic studies of fraud investigation

2.1. Method

Ethnography involves the immersive study of work practices in realistic contexts, in which the observer works within the system under study for extended periods of time, observing and documenting everyday activities as well as exceptional events. We conducted four studies, varying between two and four weeks in duration. The locations of the studies included claims handling offices of four insurance companies (both commercial and personal claims, each handling between 35,000 and 113,000 claims per annum, of which between 4% and 9% were judged as likely to be fraudulent), three specialist claims investigation units, and two investigations departments of loss adjusters (third party specialists hired by insurance companies to collect and assess evidence regarding claims and to evaluate loss for the claimant). The specialist investigations units and loss adjusters handled a smaller volume of referred claims, varying between 125 and 342 claims per annum. In each study, the researcher collected field notes of work activities in company offices, recorded meetings, interviewed company employees and inspected company documentation, formal practices and existing technologies. She also went with loss adjusters on external visits to claimants. An open and inclusive approach to data collection was adopted with all activities treated as targets for recording.

2.2. Results and discussion

We structure the report of our findings as follows. First, we summarise our observations of claims handling staff and the environment in which they operate, focusing primarily on findings that had implications for understanding how best to support fraud detection practices through technology. A more general description of the organisation, environment and nature of claims handling and fraud detection activities is presented elsewhere (Morley et al., 2006). Second, we report our observations of specialist fraud investigators and loss adjusters, overviewing the characteristics of their investigative practices that informed our systems design. Third, we describe in detail one particularly important observation that arose from our research which we harnessed for the development of technological support, that is, the central role of 'anomaly identification' in fraud detection, whereby claims handlers and investigators seem uniquely sensitised to identifying inconsistencies in data that promote suspicions as to a claim's integrity.

2.2.1 Characterising fraud detection practices

A company's potential for fraud-detection success often lies with the front-line claims handling staff who are the first members of the company to interact with the client. Yet in the companies that we examined the claims handling staff were typically distinguished from other insurance staff by their inexperience of the insurance industry. High staff turnover meant that claims handlers often had little more than six months work experience. Part of their inexperience manifested itself in a lack of fraud awareness with few staff reporting any previous knowledge of suspicious or fraudulent cases. The difficulty was compounded because high turnover rates made extensive fraud training impossible.

At an organisational level, emphasis on speedy and efficient throughput of claims meant that claims handlers had little opportunity to reflect upon individual cases, often dealing with in excess of forty cases a day. Emphasis on monitoring staff performance against these criteria, by measuring volume of claims handled, provided little incentive for identifying and referring suspect cases.

Despite their inexperience and a lack of incentive to detect fraud, claims handlers seemed readily able to spot anomalies in claims information, which is clearly a useful skill to capitalise upon. For example, one claims handler raised a concern that the insured had previously stated that their spare car keys were lost, yet had later sent them to the insurance company upon request. While not in itself an indication of fraud, it is enough to raise suspicion (e.g., that the insured had possibly lent the spare keys to a third party to stage the theft of the vehicle). The difficulty was that this, and similar anomalies, were often subsequently dismissed by inexperienced claims handlers. It seemed that claims handlers often lacked awareness and knowledge of what these anomalies might indicate. Only when a claims handler had the necessary background knowledge (e.g., local knowledge of the area in which the incident occurred) did they appear to be able to develop a suspicion about a case. Successful suspicion development was still no guarantee that a case would be detected as fraudulent as often the claims handlers would not communicate their suspicions to the loss adjusters or investigators when they referred a claim for investigation.

External sources (e.g., engineers, brokers and garages) were the main mechanism by which the front-line claims handling staff were able to identify suspect cases. For example, an engineer's report might specify that the observed damage did not match the accident circumstances. Successful interpretation and documentation of these anomalies between the claims handler and the investigators was again key to fraud detection success. We observed instances where claims handlers failed to act upon or document suspicions provided by external agents. For example, an engineer's report was sent in to the insurance company suggesting suspect behaviour by a garage. There was also a note on an engineer's report saying there was a problem with the insured's mobile number and some concerns over the validity of their purchase invoice. In both instances the claims handler failed to follow up these concerns.

To increase the chance of detecting fraudulent claims one company we observed provided the claims handling staff with a list of fraud indicators against which to check incoming claims. A fraud indicator is a factor believed to be indicative of fraud based on company experience. For each new claim, the claims handler was required to place a tick against any fraud indicator present in the claims information. Claims that demonstrate several fraud indicators were then referred for further investigation. During our observations few suspect claims were identified through this process, although the company reported financial savings attributable to the introduction of this approach. The problem appeared to be that these fraud indicators were static and therefore did not reflect the fact that the nature of fraud changes rapidly, such that as one scam gets detected another gets developed. This problem was compounded by issues of speed and efficiency in processing the main claim. This meant that the fraud indicator list was often completed some time after the initial claim call was taken and was often completed based on minimal initial claims information. Furthermore, the indicator list was typically not updated when subsequent, relevant claims information was made available.

In summary, early anomaly detection and information from external sources afford useful opportunities for the insurance industry to build on, but their success depends on suspicion development and communication of these suspicions to investigators. Fraud indicators have potential but we would need to integrate them into the claims process and ensure that they are dynamically updated.

2.2.2. Characterising fraud investigation expertise

Studies of insurance investigators revealed a wealth of specialist knowledge about common types of fraud or insurance scams. For example, an investigator built the following explanation in response to a query from a claims handler about a claim on a car that has been stolen abroad:

> "…So his friend takes the vehicle out to Spain, that may be subject to hire-purchase. Gets it out there doesn't want to take it back, throws it in as a debt, and then our man may well have found out its true pedigree and thinks, 'How am I going to get out of this one - it's subject to hire purchase?...'"

Much of this previous scenario (e.g., the notion that the car was subject to a hire-purchase arrangement) is not information that was given to the investigator in the case notes, but instead came from efforts to build a plausible explanation of the available evidence.

In addition, investigators employed rigorous strategies for testing suspicious claims that are contrary to the biases in human reasoning processes that have been observed in laboratory studies. For example, it is generally recognised that people adopt 'satisficing' approaches in hypothesis testing, in which they hold a single hypothesis at a time, considering others only when one becomes unsustainable (e.g., Ball, Maskill, & Ormerod, 1998; Evans, 2006; Simon, 1981). In contrast, the following excerpt derived from our fieldnotes illustrates how the investigator sets out to test in parallel two competing hypotheses about missing cheques, rather than focusing purely on a single salient hypothesis:

> "…Member of staff intercepted cheque; tests whether staff can associate themselves with a claim and have a cheque redirected to themselves. Supplier intercepts cheques; interviews supplier…"

Another general finding from the psychological literature is that human reasoning displays 'confirmation bias', that is, a tendency to select evidence that might demonstrate the truth of a hypothesis, rather than trying to uncover evidence that could potentially falsify the hypothesis (e.g., Nickerson, 1998; Wason, 1960). Instead of this propensity toward a confirmatory approach to reasoning, however, our observations revealed many examples of investigators adopting a falsifying or disconfirmatory stance. Often this involved a search for so-called 'disabling conditions', that is, situations that might modify the inclusiveness of a belief. This is illustrated by another excerpt from our fieldnotes, where the investigator was clarifying the conditions under which a claim might not be paid:

> "…If the company has ceased to trade you can refuse to pay the invoice… If the part of the company submitting the invoice has ceased to trade and there are no other parts of the company trading you can refuse to pay the invoice…"

In this excerpt, the investigator uses expert knowledge to point to the possibility that commercial fraudsters may trade through an associate company even when they have gone out of business. In both the pursuit of multiple alternative hypotheses in parallel, and in seeking falsifying evidence through the identification of disabling conditions, investigators demonstrate a sophistication of reasoning that is generally seen as prescriptively optimal yet is rare in many studies of human reasoning performance (e.g., Evans, 1989, 2007; see also Gale & Ball, 2006, 2009).

Our ethnographic data have proved to be immensely rich, and have yielded some important implications for tool design. For instance, tools should support the documentation and tracking of anomalies, so that the intuitive skills of even inexperienced claims handlers in detecting suspicious claims information can be capitalised upon. At the same time, claims handling needs to be conducted at speed, such that fraud detection does not impede claims through-flow. Investigators have specialist knowledge and skills, but are overloaded with cases to investigate. As such, we propose that tools are needed that can push investigation expertise closer to the front line of claims handling. In particular, the tools need to capture expertise about suspicion building, along with associated information about how suspicions might best be tested. For example, the optimal testing of certain suspicions might require the use of the 'Claims and Underwriting Exchange' (an industry-wide database of previous claims) to check for serial claimants or people that have not declared previous claims. The tools also need to encourage the pursuit of multiple alternative hypotheses and the search for disabling conditions to act as falsifying instances for putative hypotheses.

2.2.3. Developing an anomaly categorisation scheme

Before overviewing the tool prototypes, we describe some further processing that we carried out on the fieldnotes and verbal accounts gleaned during the ethnography. The aim of this processing was to elicit a scheme for categorising anomalies that we could then implement in our software design.

As Ball and Ormerod (e.g., 2000b) argue, using ethnography for systems design requires departures from a 'pure' ethnographic stance. This arises because in systems design one is observing work practices with a clear intentionality: to develop sympathetic support tools and environments. The need to convince clients to invest in technology creates additional requirements for ethnography, specifically that the data are verifiable and that they are collected in a purposive manner. The cognitive ethnographic approach adopted by Ball and Ormerod (e.g., 2000b; see also Ormerod & Ball, 2007) that was deployed in the current project also uses post-ethnographic triangulation of data across multiple methods (e.g., experimentation and verbal protocol analysis) so as to verify and make specific use of ethnographic data.

The particular post-ethnographic processing used here consisted of sorting the ethnographic observations to derive a scheme for categorising anomalies. Sorting provides a valuable method for extracting categorical structures in expert knowledge sets (e.g., Hoffman, Shadbolt, Burton, & Klein, 1995, Ormerod, Fritz, & Ridgway, 1999; Rugg & McGeorge, 1997). The aim of the sorting study was to scope the nature of anomalies in the data to see whether a scheme might be elicited that allowed claims handlers to capture the basic properties of the anomaly they observed without engaging in a lengthy dialogue. Moreover, we used the study to estimate how fast the set of classes of anomalies might be expected to grow as new cases come to light. If tools were developed that captured anomalies observed by claims handlers, then they would rapidly lose any predictive validity if the pool of anomaly types grew in an unstructured and linear fashion (i.e., if every one was unique).

Three activities were undertaken to classify anomalies. First, the researchers trawled the fieldnotes and other documentation assembled during the studies to identify all the excerpts that related in some way to suspicions or uncertainties. The resulting excerpts comprised around a fifth of the ethnographic data. Second, 50% of the excerpts were sorted using a free sort approach by the third author. This involved repeatedly sorting excerpts (each written on a separate piece of paper) into piles and sub-piles to seek a minimal set of categories under which items could be differentiated. Third, in order to estimate how the resulting multi-level scheme expanded to include new items, the remaining 50% of excerpts were sorted under the scheme, which was revised as necessary.

The trawl of ethnographic data yielded 283 excerpts. Three kinds of discrete excerpts were identified: anomalies, general beliefs and proposed types of action. Examples of excerpts showing general beliefs include: “The higher up the social ladder, the bigger the fraud is likely to be”, and “With staged accidents, the insured involves a solicitor early on”. Examples of actions include “Look out for comprehensive cover on third party vehicles”, and “Check insured hasn’t had a policy with us cancelled previously”. Further sorts took place on half the data items and revealed stable categories that mapped fairly closely onto the topic or referent of the excerpt. These categories were: insured, incident, third party, vehicle, claims handling, policy and general (see Table 1 for examples). Each category was subjected to further sorting to reveal between four and eight sub-categories. Specific examples of anomalies that were identified by the application of this sub-categorization scheme are also presented in Table 1.

(Insert Table 1 about here)

To validate our categories the remaining data items were sorted. The sort produced no further categories at the highest two levels and only two more categories at the lowest level. This suggested that our scheme was relatively stable and suitable for use in our software.

3. Development of fraud prevention software

In this section we outline the rationale and design of two software prototypes whose development was informed by the ethnographic studies: The Mass Detection Tool (MDT) that was targeted at the front-line claims handling process, and the Suspicion Building Tool (SBT) that was designed to support the processes of investigating and repudiating claims. As noted above, using ethnography for systems design departs from a traditional ethnographic stance because one is observing work practices with a clear goal in mind, which is to develop effective support tools. To do this, data must be processed to a form that can be used to specify design requirements (see Ball & Ormerod, 2000b). As part of this processing, we produced a hierarchical task analysis (e.g., Ormerod, 2000; Shepherd, 2001) of the tasks underlying claims handling. We cross-referenced and exemplified the hierarchy of tasks and plans against the ethnographic data. The hierarchy was then used to drive technical meetings with the collaborating partners.

In the original application, we proposed tools for documentation (to record searches across information repositories), visualisation (to provide graphical representations of anomalies) and simulation (to enable comparisons between known and new case patterns). The empirical research identified that, while the *functions* inherent in these proposals were useful, the idea of separate tools did not properly reflect the *tasks* of claims handling and investigation. Instead, we specified an integrated toolset structure (illustrated in Figure 1) that embodies the same functions within a two-stage process. As a by-product of implementing these tasks, the toolset offers two additional benefits: an audit trail for claims, and a repository for retaining organisational knowledge of fraud handling that companies currently lack.

(Insert Figure 1 about here)

3.1. The Mass Detection Tool (MDT)

3.1.1. Rationale for the MDT

The aim of the MDT was to act as a filter for all claims, providing the claims handler with the confidence to pay genuine claims quickly while selecting out suspicious claims for further investigation. The MDT steps the user through a question-and-answer dialogue to elicit ratings of a claim against known fraud indicators. Some indicators are supplied automatically from claims information (e.g., the car registration date) while others are informed by the claims handler's opinion (e.g., the attitude of the insured on the phone). As the indicators are rated, the system provides real-time feedback as to the probability that the current claim is fraudulent, and the probability that the claim is associated with a specific type of fraud (staged accident, inflated claim, non-disclosure, etc.).

The design of the MDT reflects a number of design decisions that emerged from the ethnography. First, it provides a structured dialogue that facilitates the communication of suspicions. Second, it provides dynamic fraud indicators whose fraud predictiveness can be updated over time. Third, in order to reduce the burden on claims handling (i.e., to limit the questions that must be addressed to those that are likely to be most fruitful) the system provides advice on the most informative indicators that have yet to be rated given the information that is currently held. Finally, the MDT provides a pull-down hierarchical menu dialogue for classifying any anomalies that the user might detect according to the anomaly coding scheme described above, guided by exemplars from a knowledge database elicited in the ethnographic studies. The intention of the anomaly registering system is to allow claims handlers to document their suspicions while minimising the documentation load and to provide a test of anomalies. If an anomaly cannot be classified then it is likely to be novel, and therefore, important, or trivial, and therefore unclassified. Either way, the procedure for unclassifiable anomalies is to refer to team managers for advice.

3.1.2. Implementation of the MDT

The core to the MDT is a Bayesian Belief Network (BBN) that calculates the probability of the current claim being fraudulent, based upon the prior probabilities of claims possessing the same range of fraud indicator ratings being proven as fraudulent. The BBN is based on a topology that links individual fraud indicators to a set of hypotheses (i.e., known fraud types) and ultimately to an overall rating for the probability of fraud. When unknown, indicators are automatically given their a priori value, which is simply how often they have been true before as a percentage of all previous claims. The calculations are based upon a store of prior probabilities of frauds given indicator ratings that is assembled from feedback of previous claim outcomes (pay, refer or refuse). So, indicators become more or less predictive of fraudulent versus genuine claims over time.

As the user changes the unknown indicator values to given ratings, the system does two things. First, it computes the impact of the new information on the current likelihood that each hypothesis is true. Second, it takes each piece of evidence that is still unknown and calculates what impact that evidence would have were it found to be true. In this way the MDT can advise the user as to which evidence is most important to investigate at any given stage. A dialectical process continues between user and system until either all information is known or the user terminates the process by deciding either to pay or to refer the claim. The MDT is non-coercive in that although it recommends whether a claim should be paid or referred, it leaves that decision to the user. This is important since the MDT's capacity to learn is partly driven by the user overriding its advice and partly by the final decision taken downstream by the fraud investigators.

Another feature of the drop-down menu classification of anomalies is that the BBN can then make use of prior probability data for each anomaly category and sub-category (derived from our ethnographic data). In this way, the system can make educated guesses about how much to update the fraud risk on the basis of anomalies associated with, say, the insured's manner, without having to know precisely the nature of the anomaly. As such, novel anomalies can affect the likelihood of a claim being referred for further investigation as they adopt the probability of the category they are most closely associated with. This enables the system to accommodate the dynamic nature of fraud.

The MDT displays active learning in three ways. First, the BBN refines its predictiveness on the basis of claims outcome information. Second, as new anomalies are entered into the system (using the Suspicion Building Tools – see following section) they also accrue predictive validity as the outcomes of claims with which they are associated change. Over time, as older fraud indicators become less predictive (e.g., as professional fraudsters become aware that insurance companies are breaking particular scams), they are replaced by newer more predictive anomalies. Third, the MDT is being reconfigured currently to allow automatic updating of the BBN topologies, using the outputs of argumentation from the investigator's tool to re-configure the topologies as older indicators are replaced by more predictive anomalies and old scams are rejected by fraudsters in favour of new ones. The latter task is by far the most complex, since it involves qualitative alterations to the BBN structure rather than simply changing the weightings of BBN nodes.

3.2. Suspicion-Building Tool (SBT)

3.2.1. Rationale for the SBT

In addition to the MDT, aimed at high-volume, front-line claims handling, we used the ethnography to specify a tool to support the processes of investigating and repudiating claims by specialist investigators and to promote best practice in these activities among less experienced staff and third parties (e.g., loss adjusters). The Suspicion Building Tool (SBT) is designed to merge automated learning capabilities with the MDT and with expert judgement in a coherent fashion, representing a form of hybrid knowledge management. For example, new anomalies are passed to the SBT for full encoding because of the need for expert review so that uninformative anomalies do not clog the system.

3.2.2. Implementation of the SBT

The SBT consists of three main components: anomaly capture and review, argumentation, and visualisation. The anomaly capture mechanism makes use of the categorisation from the MDT, but also elicits a detailed representation of anomaly terms. This representation involves the system user locating (again via drop-down menus) the precise relation between the two data items that lie at the heart of the anomaly. On a future occasion if a system user selects the same relation between two data items that are associated with an anomaly then the system will begin to collect probability data for the anomaly. Eventually, predictive anomalies come to replace previous fraud indicators, as described in the MDT section.

The core of the SBT is a rule-based argumentation engine that calculates the strength of alternative hypotheses about a claim (cf. Fox & Das, 2000). Users submit positive and negative ratings for supporting and refuting arguments to calculate the strength of these alternative hypotheses. The alternatives are initially seeded by fraud hypotheses (including the hypothesis of genuine claim) as supplied by the MDT, but deliberately without probabilities attached. As the arguments unfold, the SBT provides tailored advice, from a rule base of elicited expert search knowledge (initially seeded from expertise accrued during the ethnography), on what kinds of tests should be applied that might reveal disabling conditions that refute hypotheses. Thus, like the MDT, the SBT provides active guidance as to best practice in investigation.

The design of the SBT reflects a number of design decisions that emerged from our ethnography. First, it focuses on building suspicions by developing hypotheses and testing multiple alternative hypotheses. Second, it focuses the user's attention on disabling conditions and how they are informative for refuting arguments. Third, it is seeded with the kind of expert knowledge identified in the ethnographic research.

The SBT also has learning capabilities. In one mode, its advice modules are updated as the value of the search advice it gives becomes apparent in claims outcomes and also in argument efficiency measures (the quicker a hypothesis is refuted, the more valuable the advice is). In another mode, typical argumentation trees are learnt and can be retrieved as circumstances dictate. Finally, the argumentation trees are used to automatically update the BBN topologies of the MDT.

3.3. Toolset evaluation

Evaluation of the tools is just beginning. We are proceeding cautiously, not wishing to put systems in place before they are proven. Yet it is important to substantiate whether the knowledge embodied within the tools provides a valid and useful reflection of expert investigative practice. Our approach to assessing the knowledge bases within the tools is fourfold.

First, we adopted a bench-testing approach to check the predictive validity of the Bayesian belief network. We initially ran the MDT on a database of 200 claims taken as a random sample from the most recent six-monthly claims log of our main industrial partner (which contains 10,000+ cases). The MDT detected and flagged as problematic six cases from this sample that had been referred to the specialist investigations unit and identified a further seven claims that had been paid without referral as being worthy of further investigation. Admittedly, a sample of 200 claims is small and contains only a minority of problematic cases. However, it provides an indicator that the MDT's behaviour is predictable and stable.

We then conducted a second bench-test of the MDT in which we configured a BBN using a second company's current list of fraud indicators. We trained it with 4500 claims from an 18-month period that had triggered one or more fraud indicators in the company's paper-based system, of which 5% had been refused. After 500 cases, the percentages of refused claims and false positives (i.e., claims that were cleared of suspicion) detected by the MDT as high risk (i.e., over 80% likelihood) was 40% and 34%, respectively. After 2000 cases the figures were 100% and 4%, respectively. The sample is biased, since all cases were flagged as anomalous, but the bench-test demonstrates the capacity of the BBN for judgement refinement over time.

Second, we submitted our classification of anomalies to expert scrutiny in the specialist investigations unit of a leading insurance company. Two specialist investigators overviewed the 283 captured anomalies, beliefs and actions, as well as the categorisation scheme itself, and gave us feedback concerning the scheme that was, in the main, positive. They did not suggest the addition or removal of any categories from the main classification, though they proposed three new sub-categories. These were: "Duration of cover" (under the policy category), "Absence of insured's documentation or statements" (under the insured category) and "Unexpected or repeated match from insured's documentation or statements" (also under the insured category).

Third, we used a subset of the anomalies detected in the ethnography as materials for an experimental study of expert-novice differences in reasoning about claims. Analysis of the data from these studies is ongoing, but we have identified differences in the judgements that experts and novices draw concerning the anomalies, suggesting that the anomalies we captured do indeed reflect expert knowledge concerning likely fraud indicators.

The final assessment of the expert knowledge that we embodied within the tools must come from in-situ experience. We are embarking upon usability evaluations and then hope to undertake further ethnographic evaluations of the tools in the workplace. The key to the tools' success is that they learn: if the anomaly and investigation knowledge that the tools contain is a poor reflection of expertise, then the knowledge should be replaced over time with information that better reflects expert knowledge of fraud.

4. Discussion

This paper reports the ongoing efforts of a multi-disciplinary team to develop pro-active solutions for tackling the major problem of insurance fraud. Our particular approach has been to capture expert best practice while observing pitfalls that prevent the successful detection of fraud and to use these observations to inform the development of a suite of computer-based tools.

The adoption of a cognitive ethnography approach to data collection (Ball & Ormerod, 2000b; Ormerod & Ball, 2007) provided a rich dataset upon which to base specifications of tools that tackle key issues in fraud management. For example, the MDT overcomes problems of communication and structuring of fraud detection, and it provides expert guidance to inexperienced claims handlers. At the same time, it capitalises upon the fact that all people, regardless of experience, are good at detecting anomalies. Likewise, the SBT provides an environment that fosters best practice in investigation, including support for testing multiple alternative hypotheses through falsification.

Our ethnographic studies have thrown up some important challenges to standard laboratory findings regarding human reasoning (for reviews see Evans, 1989, 2007). In this regard, we believe that an ethnographic stance offers valuable insights into realistic work practices that simply cannot be gained by other methods. For example, traditional laboratory studies have demonstrated two key limitations on human hypothesis generation and testing, namely a tendency to 'satisfice' by focusing on a single hypothesis rather than considering alternatives (e.g., Ball, Maskill, & Ormerod, 1998; Evans, 2006; Simon, 1981), and a tendency towards 'confirmation bias' (e.g., Nickerson, 1998; Wason, 1960). Our studies of fraud investigation expertise reveal neither of these limitations: expert investigators typically explored multiple hypotheses concurrently and they selected evidence that would falsify as well as verify their hypotheses. The absence of confirmation bias follows necessarily from the pursuit of multiple alternative hypotheses, since evidence that might verify one hypothesis will serve as a falsifying case for alternative hypotheses (for related arguments and findings see Gale & Ball, 2006, 2009; Gorman, 1995; Klahr & Dunbar, 1988; Okada & Simon, 1997). Thus, ethnographic studies such as the present one can have important messages for systems designers and can contribute valuable insights to psychological theorising.

The tools themselves demonstrate an interesting hybrid approach to inference engine development, combining the probabilistic inference derivation of BBNs with the rule-based derivation of inferences through argumentation. Having two types of inferential mechanisms proved necessary because the tasks being tackled by the MDT and the SBT differ: the MDT simply prioritises claims for investigation according to prior probabilities for fraud, while the SBT provides a support environment for evaluating competing hypotheses. All too often, the first thing that is fixed in developing an intelligent software system is the choice of underlying processing architecture (e.g., rule-based, fuzzy logic, Bayesian, neural net). The research shows the importance of informing this choice by a thorough task analysis and empirical study of the domain in which an intelligent system is to be applied. It is only by identifying the actual tasks that must be accomplished by the human users of intelligent support tools that the best choice(s) of underlying processing architecture can be made.

Acknowledgements

This research was supported by a grant from the DTI/EPSRC Management of Information initiative, no. L1243234322. We thank all the companies that took part in the studies and our other project partners for their assistance.

References

Association of British Insurers, 2006. *Submission to the fraud review* [online]. Association of British Insurers, UK. Available from: http://www.abi.org.uk/Bookshop/default.asp#Fraud_Prevention [Accessed 11 June 2008].

Anderson, R.J., 1994. Representations and requirements: The value of ethnography in system design. *Human-Computer Interaction, 9,* 151-182.

Ball, L.J., Maskill, L., & Ormerod, T.C., 1998. Satisficing in engineering design: Causes, consequences and implications for design support. *Automation in Construction, 7,* 213-227.

Ball, L.J., & Ormerod, T.C., 2000a. Applying ethnography in the analysis and support of expertise in engineering design. *Design Studies, 21,* 403-421.

Ball, L.J., & Ormerod, T.C., 2000b. Putting ethnography to work: The case for the cognitive ethnography of design. *International Journal of Human-Computer Studies, 53,* 147-168.

Button, G., 2000. The ethnographic tradition and design. *Design Studies, 21,* 319-332.

Chan, P.K., Fan, W., Prodromidis, A.L., & Stolfo, S.J., 1999. Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems, 14,* 67-74.

Cox, K.C., Eick, S.G., Willis, G.J., & Brachman, R.J., 1997. Visual data mining: Recognizing telephone calling fraud. *Data Mining and Knowledge Discovery, 1,* 225-231.

Crabtree, A., Nichols, D.M., O'Brien, J., Rouncefield, M., & Twidale, M.B., 2000. Ethnomethodologically informed ethnography and information system design. *Journal of the American Society for Information Science, 51,* 666-682.

Evans, J.St.B.T., 1989. *Bias in human reasoning: Causes and consequences.* Hove, UK: Lawrence Erlbaum Associates.

Evans, J.St.B.T., 2006. The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review, 13,* 378-395.

Evans, J.St.B.T., 2007. *Hypothetical thinking: Dual processes in reasoning and judgement.* Hove, UK: Psychology Press.

Experian Decision Analytics, 2008. *How to prevent fraud* [online]. Experian Decision Analytics, UK. Available from: http://www.experian-da.com/Web/Solutions/CA/Fraud.html [Accessed 11 June 2008].

Fawcett, T., & Provost, F., 1997. Adaptive fraud detection. *Data Mining and Knowledge Discovery, 1,* 291-316.

Fox, J., & Das, S., 2000. *Safe and sound: AI in hazardous applications.* California. AAAI Press.

Gale, M., & Ball, L.J., 2006. Dual-goal facilitation in Wason's 2-4-6 task: What mediates successful rule discovery? *Quarterly Journal of Experimental Psychology*, *59*, 873-885.

Gale, M., & Ball, L.J., 2009. Exploring the determinants of dual goal facilitation in a rule discovery task. *Thinking and Reasoning, 15,* 294-315.

Gorman, M.E. 1995. Confirmation, disconfirmation, and invention: The case of Alexander Graham Bell and the telephone. *Thinking & Reasoning, 1*, 31-53.

Hammersley, M., & Atkinson. P., 1995. *Ethnography: Principles in practice (Second edition).* London: Routledge.

Hoffman, R.R., Shadbolt, N.R., Burton, A.M., & Klein, G., 1995. Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes, 62,* 129-158.

Hollien, H., Geison, L., & Hicks Jr, J.W., 1987. Voice stress evaluators and lie detection. *Journal of Forensic Sciences, 32*, 405-418.

Horvath, F., 1982. Detecting deception: The promise and the reality of voice stress analysis. *Journal of Forensic Sciences, 27,* 340-351.

Hughes, J.A., Sommerville, I., & Bentley, R., 1993. Designing with ethnography: Making work visible. *Interacting with Computers, 5,* 239-253.

i2 Solutions, 2008. *Insurance fraud.* I2 Solutions, UK. Available from: http://www.i2.co.uk/Solutions/InsuranceFraud/default.asp [Accessed 11 June 2008].

Klahr, D., & Dunbar, K., 1988. Dual space search during scientific reasoning. *Cognitive Science, 12*, 1-48.

Morley, N.J., Ball, L.J., & Ormerod, T.C., 2006. How the detection of insurance fraud succeeds and fails. *Psychology, Crime, and Law, 12*, 163-180.

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

Okada, T., & Simon, H.A., 1997. Collaborative discovery in a scientific domain. *Cognitive Science, 21*, 109-146.

Ormerod, T.C., 2000. Using task analysis as a primary design method: The SGT approach. *In* J.M.C. Schraagen, S.F. Chipman, & V.L. Shalin, eds. *Cognitive task analysis*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ormerod, T.C., & Ball, L.J., 2007. Qualitative methods in cognitive psychology. *In* C. Willig & W. Stainton-Rogers, eds. *Handbook of qualitative research in psychology*. London: Sage Publications Ltd., 553-574

Ormerod, T.C., Barratt, E.C., & Taylor, P.J., 2008. Investigative sense-making in criminal contexts. *In* J.M.C. Schraagen, L. Militello, T. Ormerod, & R. Lipshitz, eds. *Macrocognition and Naturalistic Decision Making.* Aldershot, UK: Ashgate Publishing Limited.

Ormerod, T.C., Fritz, C.O., & Ridgway, J., 1999. From deep to superficial categorisation with increasing expertise. *In* M. Hahn, & S.C. Stoness, eds. *Proceedings of the 21st Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates, 502-506.

Rugg, G., & McGeorge, P., 1997. The sorting techniques: A tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems, 14,* 80-93.

Shepherd, A., 2001. *Hierarchical task analysis*. London: Taylor and Francis Inc.

Simon, H.A., 1981. *The sciences of the artificial (Second edition).* Cambridge MA: MIT Press.

Wason, P.C., 1960. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129-140.

Yue, D., Wu, X., Wang, Y., Li, Y., & Chu, C-H., 2007. A review of data mining-based financial fraud detection research. In *Proceedings of Wireless Communications, Networking and Mobile Computing—WiCom 2007*. New York: Wiley-IEEE Press, 5519-5522.

Table 1.

Full categorisation scheme for anomalies observed in the ethnographic data

Category	Sub-categories for 'insured'	Examples
Insured	Previous claims history	"A recent small claim followed by a larger claim can be evidence of toe dipping"
	Financial background	"Higher up the ladder the bigger the fraud"
	Insured's manner	"People are not always good on the phone and are not necessarily trying to be obstructive"
	Personal characteristics	"Old people don't drive high value fast cars"
	Suspicious behaviour	"He has sent keys and documents in straightaway - you wouldn't know to do this"
	Mismatch from insured's statements	"Time of accident originally stated as 9.30 then changed to 9.25"
	Appears too genuine	"Look out for straightforward circumstance, for example, where the insured has gone into the back of the third party"

Category	Sub-categories for 'incident'	Examples
Incident	Damage/injury does not match reported incident	"Engineer says can't be vandalism as bumper could not have been pulled off by a person"
	Fraud hotspot	"There are hotspot postcodes: BB/LU/HW"
	No witnesses or independent evidence	"Staged accidents ...no witnesses, car immediately taken from the scene, no police involvement"
	Site of incident (secure/impossible/suspect)	"Believes the car was parked on a street with double yellows"
	Stereotypical easy fraud type	"Suspicious claims would be something like nail varnish on the carpet"
	Surprising coincidence	"Two claims from the same company on the same date both with excessive windscreen claims"
	Time of day	" Staged accidents occur early morning, late evening"

Category	Sub-categories for 'third party'	Examples
Third party	Insured/TP relationship	"The family member who was interviewed admitted TP knew the insured"
	Missing TP data	" Suspicious of the fact that the insured says they hit TP yet no TP details on claim form at all"
	Claims history of TP	"The passengers duplicated the previous claims"
	Number of TPs	"The insured admits to hitting two TPs yet they have correspondence from four TPs"
	TP claim mismatches incident description	" Anomaly of no damage to either vehicle, the insured has a baby in the back who didn't even wake up, and the TP has whiplash and psychological distress"

Category	Sub-categories for 'vehicle'	Examples
Vehicle	Insufficient or mismatching identity information	"The vehicle is a different colour to that stated on policy"
	Low value/old vehicle	"C-reg Sierra was used as a courtesy car; FC says he probably just borrowed it off his mate"
	Missing component	"Steering wheel missing when engineer inspected vehicle; could be it was a funky steering wheel and it's a cherished car and he's taken it off for himself"
	Unrecovered/burnt out	"Vehicle not available for inspection"
	Unexpected/repeated match from insured's documentation/statements	"Suspicious of lots of previous searches done on a car when looking it up on Goldcar [a Spanish car hire company]"
	Value claimed too high/low	"Discrepancy between the insured's valuation and the price of the vehicle"

Category	Sub-categories for 'claims handling'	Examples
Claims handling	Dubious claims handling parties	"A dodgy solicitor has been identified in the claim"
	Excessive claim	"Disparity between reserve put on claim for personal injuries and the number of passengers"
	Mismatch from engineer's documentation	"Originally claimed for 6 vehicles set fire to but there are now 7 engineers reports"
	Suspicious invoice	"Invoice company has ceased to trade according to Experian"
	Timing anomalies relating to notification	"Hire period of vehicle starts two days after accident"

Category	Sub-categories for 'policy'	Examples
Policy	Claimant not covered under policy	"Guy took insurance out in ex-partner's name to take advantage of no claims bonus - discovered when he damaged courtesy car and they rang partner"
	Excessive cover	"It is suspicious when premium for comprehensive cover is more than the value of the vehicle"

Category	Sub-categories for 'general'	Examples
General	General	"Believes the insured is trying to get one over the wicked insurance company"

Figure 1.

Overall structure of the FRISC toolset: Hexagons are inference engines, boxes are system/user dialogues, and arrows show feed-forward between system elements.