Abstract—Credit card fraud analysis is almost entirely automated. However, there may be occasions when a human analyst is required to intervene. In this paper, we consider situations in which a transaction triggers an automated alert but not sufficiently to allow automated response. On such occasions, automated analysis makes a recommendation as to the fraud pattern that has been identified and the human analyst decides, if this recommendation is correct and what action to take. In order to support the analyst, a “dashboard” can be used to display information that is relevant to the fraud pattern. Thus, a computer could analyze transaction data, define this as a known fraud pattern, and then present the analyst with a dashboard to illustrate how the transaction might fit the pattern. We explore the efficiency with which people respond to the computer’s recommendation, and whether computer confidence has an impact on this response. We define efficiency in terms of information search: the user will either have the relevant information on screen or will need to drill-down into a dashboard (i.e., open additional windows for information). The results show that participants adapt their decision making to the confidence of the automated support, and, although they drill-down even when not required, are efficient in terms of time spent looking at relevant information.

Index Terms—Automation transparency, credit card fraud, decision support systems, human-automation interaction.

I. INTRODUCTION

The analysis of credit card fraud has become highly automated in recent years [5], [8], [9], [13], [30]. However, there remain situations in which a human analyst might be required to contact a cardholder to check a transaction or to review the decisions made by automated systems. For example, in the EU-funded SPEEDD project, online inductive logic programming (using Online Learning of Event Definitions (OLED)) is applied to credit card transaction data in order to learn relational patterns in fraudulent activity [1]. In this project, the set of fraud patterns is as follows:

1) increasing amounts of money spent on a single card over a sequence of transactions (increasing amount—IA);
2) using a card to make a very large transaction, compared to the average transaction in the region (large amount—LA);
3) a very high number of transactions in a very short time period (flash attack—FA);
4) transactions in geographically remote places in a short time period (transactions in faraway places—TF).

In credit card fraud investigation, automation analyzes and flags suspicious transactions. Ideally, the automated system would process a transaction and make a decision in milliseconds. When automation confidence (or, certainty) is very high, transactions are labeled as fraud and cards are automatically blocked. When there is no indication of fraud, the transaction is allowed. However, there are situations where there is some indication of fraud but not enough evidence to take an automated action, e.g., there might be a new form of fraud, or the data do not quite meet the threshold required [13]. In such cases, the analyst might review the computer’s recommendation and then decide whether to call the cardholder to obtain additional information, or permit, or block the transaction. Visualizations of information relevant to the transaction can aid the analyst. Such visualizations can be designed on an analogy with the “dashboard” of an automobile: “When properly designed… dashboards support a level of awareness… that could never be stitched together from traditional reports” [11, p. 5]. A dashboard offers a high-level perspective on the situation, with the opportunity to drill-down to lower levels of detail, such that interaction follows discrete stages, e.g., “Overview first, zoom and filter, then details on demand” [31] or “Analyze first, show the important, zoom/filter, analyze further, details on demand” [15].

Some fraud patterns require contextual information for proper diagnosis, while others require the investigation of transaction-specific “detailed” information. For fraud analysis, contextual information could include a transaction suspiciousness score, which could be a probability between 0 and 1 based on the bespoke algorithm of a particular organization, together with the

1https://github.com/ishkin/Proton/
2The dataset used for this paper can be publicly accessed from the project website http://speedd-project.eu/data
type of transaction made, e.g., cardholder present (in a store) or cardholder not present (online), time of day, geographical region of transaction, customer profile, and spending behavior [7], [22], [29], [30]. It is not clear whether it is more important to show the context in which a suspected fraudulent transaction has occurred (i.e., overview) or show the specific flagged transaction (i.e., details), as different fraud types are diagnosed using different types of information [8], [9], [29]. Moreover, it is not known how different modes of interaction affect the analyst’s information search. People tend to optimize information search, depending on the task they are required to perform and on information accessing costs, so that relevant information gain per unit time is maximized [6]. From the concept of “information foraging” [25] we assume that people filter out irrelevant information sources and select relevant ones based on previous searches. By analogy, the specific information sources used by the automation could be shown to the human analysts and would have (because they are on the display) a low access cost. Other information could be accessed through pop-up windows, which would have higher access costs. The four fraud types outlined above would place different weight on the relevance of available information. In this study, we investigate whether presenting participants with information relating to these specific fraud types, first, decreases their decision time and, second, changes their information search behavior (in terms of drill-down activity), and third, whether the confidence of the automation influences these measures.

II. CREDIT CARD FRAUD ANALYSIS

The credit card industry is understandably very protective of the approaches used in the analysis of credit card fraud. While we have benefitted from discussions with a number of fraud analysts operating in the U.K., Europe, and the USA, the following description presents a high-level account of decision making in which analysts engage. This does not represent analysis conducted by any individual organization, but a general description of how analysis is approached. The description of this process underpins our experimental design.

Fig. 1 shows a decision process for credit card fraud analysis. In terms of the system output, a transaction suspiciousness or risk score will be based on probabilities that are defined in terms of an organization’s risk models. The risk model will be tailored for specific types of client, region, transaction, etc., but could include such measures as number of transactions for an account in a given time period, value of a transaction, number of cash withdrawals at automated teller machines, etc. The risk models inform the design and operation of algorithms used by the automated system.

Credit card fraud can involve several types of analysts, from those involved in the definition and running of machine learning algorithms on big datasets to those who respond to alerts from the fraud detection system. In cases involving response to alerts, call-center analysts will triage these according to risk level; alerts with the highest level of risk are worked on first (i.e., “priority mode”). From our interviews, we believe that credit card organizations can employ between 250 and 1500 fraud analysts, and the total number of cases to be processed by an analyst is around 200 per day. In terms of the baseline decision time, a typical decision by a call handler might take around 1.8 min (assuming 200 cases to investigate in a 7.5 h working day, with 1.5 h of breaks). Given that the handling of a case involves blocking or allowing the transaction (and completion of forms for audit purpose) or speaking to the cardholder prior to making the decision (and so involve a telephone call as well as form filling), one might anticipate the decision on the fraud (based on the information provided) to be performed relatively quickly. Consequently, having a system that supports quick but accurate decision making could be advantageous. Thus, a well-designed dashboard, together with a reliable recommender system, ought to minimize time spent on each call.

If a transaction meets the criteria to which the algorithms apply, then it would be automatically blocked. However, if only some of the criteria were met or if there was some uncertainty concerning the criteria, the transaction might be presented to a human analyst. In this paper, we are interested in the role of call-center analysts, who perform customer verification on suspicious transactions. In general, the decision to contact a customer would be made if the automated system was able to match some but not all of its criteria. In this instance, the call-center analyst would be presented with some of the transaction details together with the automated output (in the form of a score). In Fig. 1, this activity would involve the top three boxes. The call-center analyst would interpret information from the dashboard in order to define the situation; if there was insufficient information, then the analyst would contact the cardholder. In the telephone call, the call-center analyst would follow a clearly defined script in order to establish whether the...
person at the other end of the phone is the genuine cardholder and whether the cardholder made the purchase or whether it was made fraudulently. The call-center analyst would then confirm that the transaction was acceptable or mark it as suspicious. This could involve reimbursing the cardholder or could trigger further investigation. For this role, analysts have access to transaction and customer details, both past and present.

III. HUMANS AND AUTOMATION

A. Recommender Systems and Transparency

Recommender systems provide suggestions for a specific action or solutions to a problem [28]. Automated reasoning on the data computes an answer and displays it to the user, e.g., in the form of a recommendation for an action to be taken, or in the form of a detected event. In the context of fraud handling, this can relate to the detection of aspects of transactions which are related to definitions of fraud (see above). “Transparency” is a defining factor of a “good” recommender system [34], i.e., the extent to which the computational process behind the recommendation is visible and clear to the human. It has been shown that increasing transparency of recommender systems, that is, making explanations available to the user along with recommendations, improves decision performance [26], [33], [34].

In this paper, we display recommendations in terms of a specific fraud that has been identified, with the computer’s confidence in its recommendation, and the most relevant data.

B. Dealing With Automation Confidence

If the automation has high confidence in its recommendation, then the role of the human could be to confirm this. A well-supported observation is that human-automation system performance with “perfect” (i.e., high reliability) automation tends to be superior to that without automation, and performance with “imperfect” (i.e., low reliability) automation tends to be inferior to that without automation [16], [20]. What makes the findings of these studies counterintuitive is that participants, in the high-reliability conditions, will follow the advice of automation even when it is wrong. This could be interpreted as automation bias, where humans are complacent and not checking the automation output against other information sources [23], [24], i.e., the human merely accepts the automation’s output and does not contribute to the decision making. Conversely, if the automation has low confidence, then the human might ignore useful output. Such automation bias has been interpreted in terms of operator characteristics, e.g., individual differences [10], experience of automation failing [4], decision accountability [33], or the design of the automation, e.g., level of automation [18], or use of information sources [21], [32].

From a review of automation reliability, Wickens and Dixon [36] conclude that there is a “cross-over” point at 70% reliability, below which unreliable automation was worse than no automation (in terms of human-automation system performance). However, operators might use unreliable automation in order to free up cognitive resources, which would explain reliance on unreliable automation under high workload [36], or operators might use the recommendation as a cue to apply their own heuristics [19].

It has been suggested that acceptance of (and, by implication, trust in) automated decision support can be considered in terms of “conformance” between human and automation problem-solving style [35]. In other words, if the automation appears to be addressing the problem in a manner that is similar to the one used by the human, then this could lead to higher level of acceptance by the human. For this paper, conformance involves the decision to block or allow a transaction on a credit card and the use of available information.

IV. DESIGNING DASHBOARDS FOR FRAUD ANALYSIS

In this experiment, dashboards are designed according to visualizations that we had observed to be common in fraud analysis; the aim was not to produce a single organization’s display but to produce designs that had sufficient family resemblance across the domain. In addition, the dashboards needed to present the confidence that the computer applies to its classification of the fraud pattern, and to present the most relevant information in support of that classification.

The experiment also considered the need to drill down for information. By “drill down,” we mean whether the participants need to call pop-up windows (see Figs. 2 and 3) to obtain further information. Some fraud types require contextual information, such as information about the normal card usage in the region where it occurred. When a transaction is outside the bounds of what is considered “normal,” this is a potential indication of fraud. For other fraud types, specific transaction information is needed for diagnosis. For example, it is sufficient to know that one card has been physically used in two distant countries in a very short amount of time to suspect that the card may have been cloned.

In this experiment, we contrast an “overview” design (which summarizes transactions occurring in a particular country) with a “detailed” design (which provides a view of transactions on a given card). In each dashboard, it was possible to identify some fraud patterns solely on the information presented to the user, but there was the option to “drill down” to find further information—which meant that all of the fraud patterns used in this experiment could be identified using both dashboard
designs. Depending on the fraud pattern, one might assume that a given dashboard design would be more suitable, and, as the use of pop-up windows is only relevant to some of the frauds, efficient performance could be defined in terms of their use. This is shown in Table I.

The overview dashboard (see Fig. 4) presents the user with a summary bar, on the top of the dashboard, which shows the total number of transactions investigated by the department, the number of transactions flagged by the automation, the average amount (of cash spent on a transaction), and the average volume (of transactions) for a selected country. To select a geographical region, the user clicks on the map on the right of Fig. 4.

The analyst also has a list of the patterns flagged by the automation in the form of an event list (on the left of Fig. 4). The event list shows the transaction number, the fraud pattern identified, the automation confidence level associated with this flag, and the current status of the pattern (not investigated, fraud, contact, allow). The analyst would take the next “not investigated” fraud from the event list and work on this.

At the bottom of each dashboard there are five buttons: three of them are for the possible decisions the user can make: allow a transaction, query the transaction with the cardholder, block the transaction. On both dashboards, pop-ups are called by clicking the “explain” buttons: the one on the map brings up information related to the selected country (see Fig. 2), and the other, under the “events list,” brings up more data related to the selected pattern (see Fig. 3).

As Table I indicates, when using the overview dashboard, analysts either have the information they require for diagnosis of fraud, or they have to search for it in the list pop-up.

In the detailed dashboard (see Fig. 5), participants can use the information available on screen or have to bring up the country pop-up window to get contextual information. The window on the left labeled patterns to investigate is the same as the event list window in the overview dashboard. At the top of the pattern view window (see Fig. 5), visualizations show the number of transactions and the interval between them (left) and transaction amounts (right). Below these, a map shows the geographical region in which the transaction(s) took place, and, at the bottom of the window, the date and time of flagging is shown along with the automation confidence level and reason for flagging. It should be noted that while Figs. 4 and 5 show different event lists, participants encountered all of the possible events during their trials.

V. EXPERIMENT

We are interested in whether the participants’ time to make a decision is affected by automation confidence. We are also interested in how participants use the information available and whether their information-search strategy (i.e., how many pop-ups participants open and how long they choose to keep them open) changes with different confidence levels and/or fraud types. These provide measures of whether users recognize the need for extra information, and the effort put into seeking additional information for the transaction. Thus, the experiment was designed to explore the following two questions.

### Table I

<table>
<thead>
<tr>
<th>Fraud Pattern</th>
<th>Information Required</th>
<th>Overview Dashboard</th>
<th>Detailed Dashboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA: increasing amounts</td>
<td>a) number of transactions</td>
<td>List pop-up required</td>
<td>List pop-up required</td>
</tr>
<tr>
<td></td>
<td>b) amounts trend</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IA: large amount</td>
<td>a) transaction cost</td>
<td>List pop-up required</td>
<td>Country pop-up required</td>
</tr>
<tr>
<td></td>
<td>b) country average amount</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FA: flash attack</td>
<td>a) number of transactions</td>
<td>List pop-up required</td>
<td>Country pop-up required</td>
</tr>
<tr>
<td></td>
<td>b) country average volume</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF: transactions in faraway places</td>
<td>a) countries where transaction were made</td>
<td>List pop-up required</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b) time between transactions</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fig. 4. Overview dashboard.

Fig. 5. Detailed dashboard.

1) Does user behavior vary with automation confidence?
2) Does information search-strategy differ when people are presented with information relevant to their decision activity?

We would consider optimal behavior to seek extra information only when it is required and to keep these pop-up windows opened until the needed information is extracted. We would also expect human decision making to reflect automation confidence levels, in line with previous work [2], [3], [19], [36], i.e., users will allow more transactions in the low confidence cases and flag more transactions as fraud in the high confidence cases, and make more decisions to contact customers in the medium confidence cases.

A. Participants

A total of 27 people took part in the experiment [17: male; 10: female; age range: 22–29]. None of the participants had experience of working in the credit card industry or financial sector. Given the observation that fraud schema are highly company-specific, we felt that it was inappropriate to recruit participants from a specific company because they would respond to the information according to their organizational policy. Having said this, the experiment has been repeated with a small number (four) of experienced credit fraud analysts, and performance is consistent with this study [1]. For our recruitment of participants, we assume that people educated to degree level and given training on
the fraud patterns to identify would provide a reasonable proxy with call-center agents, whose role is primarily to follow the script provided to them. Therefore, each participant was trained to criterion (see below) before the experiment began. In this way, we have a homogenous user group from which to explore the impact of dashboard designs on decision performance.

B. Procedure

The study was approved by the University of Birmingham Ethics Panel (Reference Number ERN_13-0997). All data were anonymized and participants provided informed consent.

Following a briefing on the task and training in using both dashboards, participants were given a demonstration of how fraud patterns could be recognized from the data presented in a dashboard and the other windows. Next, they were given four practice trials (to become accustomed to interacting with the dashboard) in order to familiarize themselves with the dashboard before beginning the trial. They were asked to process up to ten examples. Once participants were able to correctly process five examples consecutively the main experiment began. In order to replicate the script-based aspect of call-center analysts’ work, participants were provided with an aide memoire, which defined the four fraud patterns.

Each participant investigated 48 patterns, 24 using each dashboard. The set of 24 patterns (for each dashboard) were randomized across participants in order to minimize order effects. The patterns were defined in terms of four fraud types—increasing amounts (IA); transactions in faraway places (TF); large amounts (LA); flash attack (FA)—and three levels of automation confidence—low (≤71%); medium (51%–69%); high (≥70%). The rationale for defining “high” confidence as > 70% (rather than, say, 90% or 100%) was twofold. First, we assumed that if confidence is sufficiently high, then the decision will be made automatically, and thus only those decisions which fall below a threshold will be passed on to the human operator. Second, a reliability of 70% could be taken as a cutoff point, below which there is a drop in performance such that it is preferable to ignore unreliable automation [31]. Each fraud pattern was presented twice under each automated confidence level.

The independent variables for the experiment were: dashboard, fraud pattern, and automation confidence level. The dependent variables were: time to submit a decision, number of pops-up opened, and decision type. All statistical tests were performed using IBM SPSS v24, and tests of normality were applied to the data in order to select statistical tests [5].

C. Results

1) Average Time to Make Decisions: Tests of normality (using Shapiro–Wilk) indicated that the average time to make a decision did not follow a normal distribution in the low and medium-confidence conditions. Thus, nonparametric tests were applied to the time to make decision data. There was a significant difference between the high confidence and the medium-confidence conditions (z = −2.407, p = 0.016) (median high = 11.73 s, median med = 13.06 s) for time to make decision. No other differences were found (see Fig. 6).

2) Type of Decision Made by User: The types of decision data [allow transaction, block transaction (i.e., fraud) or contact cardholder] were normally distributed except for fraud in the low-confidence condition and allow in the high-confidence condition. In cases, where normally distributed datasets were compared, parametric tests were applied otherwise nonparametric tests were applied. When automation confidence level was low, there were significantly more decisions to allow than block the transaction (z = −3.69, p < 0.001) (median allow low = 56%, median fraud low = 13%). There were also more decisions to contact the customer in the low condition than to flag the transaction as fraud (i.e., block it) (z = −3.546, p < 0.001) (median contact low = 31%, median fraud low = 13%).

In the medium computer confidence cases, the number of decisions to contact the customer was significantly higher than to allow transactions (t(24) = 2.923, p = 0.007) (mean contact med = 42.8%, mean allow med = 24.12%).

When automation confidence level was high, participants were significantly more likely to mark the transaction as fraud than to either allow (z = −3.919, p < 0.001) (median fraud high = 56.25%, median allow high = 0%) or contact the customer (t(24) = 3.131, p = 0.005) (mean fraud high = 59.5%, mean contact high = 32.25%). Moreover, there were more decisions to contact the customer than to allow transactions (z = −3.546, p < 0.001) (median contact high = 31.25%, median allow high = 0%).

Furthermore, the proportion of allow decisions in the low-confidence condition was higher than in the medium-confidence condition (t(24) = 5.85, p < 0.001) (mean allow low = 52.36%, mean allow med = 24.12%), which was higher than in the high-confidence condition (z = −3.472, p = 0.001) (median allow med = 25%, median allow high = 0%). Fraud decisions are higher in high-confidence condition than the medium (t(24)
Fig. 7. Types of decision made by the user, for both dashboards, under different automation confidence levels.

Fig. 8. Average time country pop-ups were kept open. Outliers are indicated as dots.

Fig. 9. Average time list pop-ups were kept open. Outliers are indicated as dots.

= 4.76, \( p < 0.001 \) (mean fraud high = 59.5%, mean fraud med = 35.32%) and from the medium to the low-confidence condition (\( z = -3.415, p = 0.001 \) (median fraud med = 38%, median fraud low = 13%). There were no differences in contact decisions between confidence levels. These can be seen in Fig. 7.

3) Pop-Ups Opened: Normality tests showed that data for the number of pop-ups opened were not normally distributed. A Wilcoxon test showed significant differences between the required and not required conditions for list pop-up and country pop-up. Participants looked at the list pop-up for longer in the required condition compared to the not required condition (\( z = -8.504, p < 0.001 \) (median list required = 6.26 s, median list not required = 3.65 s). However, participants looked at the country pop-up for longer when it was not required (\( z = -2.306, p = 0.021 \) (median country required = median country not required = 0 s). The list pop-up was kept open for longer than the country pop-up both when required (\( z = -10.687, p < 0.001 \) (median list required = 6.26 s, median country required = 0 s) and when not required (\( z = -12.608, p < 0.001 \) (median list not required = 3.65 s, median country not required = 0 s).

4) Time Pop-Ups Active: The Shapiro–Wilk test showed that data for the time pop-ups were kept active were not normally distributed. A Wilcoxon test showed significant differences between the required and not required conditions for list pop-up and country pop-up. Participants looked at the list pop-up for longer in the required condition compared to the not required condition (\( z = -8.504, p < 0.001 \) (median list required = 6.26 s, median list not required = 3.65 s). However, participants looked at the country pop-up for longer when it was not required (\( z = -2.306, p = 0.021 \) (median country required = median country not required = 0 s). The list pop-up was kept open for longer than the country pop-up both when required (\( z = -10.687, p < 0.001 \) (median list required = 6.26 s, median country required = 0 s) and when not required (\( z = -12.608, p < 0.001 \) (median list not required = 3.65 s, median country not required = 0 s).

VI. DISCUSSION

The results are considered in terms of the questions posed in Section V.

1) Does user behavior vary with automation confidence level?

Participants responded faster to transactions associated with a high automation confidence level compared to the medium confidence transactions (although there was no significant difference between low confidence and either medium or high). Automation confidence level had a bearing on the type of decision that participants made: the greater the automation confidence, the more likely participants were to block the transaction, and the lower the automation confidence the more likely participants were to allow the transaction. We suggest that
this points to two reasonable assumptions that the participants could be making: if the computer has made an error, it might not have useful information available to it, and speaking to the cardholder would involve a sort of unstructured query (which can be refined during the conversation, albeit on the basis of a script). In other words, when automation confidence is low, then it makes sense to seek information outside the computer before deciding to declare a transaction as fraudulent, but if the computer confidence is high (but you are not sure), then it makes sense to speak to the cardholder before allowing a transaction. In the medium-confidence conditions, then speaking to the cardholder was the most commonly selected option and this makes sense because it reserves judgment on the available options. This makes sense because it reserves judgment on the available options.

2) Does information search-strategy differ when people are presented with information relevant to their decision activity?

In all conditions, participants used at least one pop-up—even when this was not required. When using the “list” pop-up, participants tended to make more use of the pop-up windows when they were required, and to keep these open when they were required, than when they were not required. This was not so apparent for the country pop-ups, and participants kept the country pop-up open for longer when it was not required. This suggests that the relevance of the information in these pop-ups differed in terms of their content. This raises some interesting questions concerning the issue of “transparency.” While the dashboards in this experiment had been designed to provide necessary information to support the recommendation for each specific fraud type, it is possible that participants did not regard this as sufficient—and so, would seek more information. This could relate to the suggestion that people seek more information to increase their confidence in a decision [12], [23], even when the additional information might not be useful. From this, the act of seeking additional information might not relate directly to the decision making activity so much as their emotional response to making a decision. It also implies that estimates of the value, or relevance, of information might differ from the actual value (in terms of a specific decision). It is apparent, in our experiment, that participants were able to make reasonable judgments about the value of information in the “list” pop-ups, but were less able to make such judgments for the map pop-up. For the list pop-ups, analysis shows that even though extra information is brought up, it is rapidly discounted when it is not required. When the information present in the list pop-up is required to make a decision, participants spend significantly longer looking at it than when it is not required. This effect does not occur in the case of the country pop-up, but this may be due to the fact that there was very little additional information in this window compared to the list pop-up and the users and so participants may have responded to them differently (see Figs. 2 and 3). It might be the case that the very fact that the country pop-up contained information that was not relevant could have evoked a longer decision time (to ensure that the information was either redundant or not relevant). Alternatively, it might be the case that what looks like a search for additional information is actually a means of bounding the information space (and determining what information to ignore). In support of this suggestion, it is worth noting that the fraud patterns we defined were not mutually exclusive (i.e., information can fit more than one pattern), and so the task of aligning information with a pattern is somewhat fuzzy. In response to this fuzziness, a reasonable response might be to rule out alternative explanations, and this could involve checking the other information that is available. Certainly, in our postexperiment debriefs, this strategy was mentioned by several participants. Thus, presenting information in an easily accessible dashboard is only part of the challenge of supporting decision making and people will respond to recommendation in terms of the transparency between their understanding of the task and the information available.

VII. CONCLUSION

We report the design of dashboards for analyzing four types of credit card fraud under different levels of computer confidence. The dashboards differed in the type of information presented to the user. Based on the information they presented and the type of fraud being investigated, efficient pop-up use was defined. Results show that participants would drill-down for information even when this is not required. However, at least in the case of the list pop-up, participants quickly discount the window when it was not required.

Relating these findings to the notion of conformance [35], it can be seen that participants adapted their decision to align with the confidence of the computer and that their use of available information was, to some extent, dependent on the type of fraud that was being flagged. In [35, p. 50], there was the suggestion that “conformance may only be relevant for expert users who hold consistent and well-developed decision-making strategies.” Our findings suggest that training participants to criterion, providing them with an aide memoir to define the task, and providing clearly defined dashboards all help to ensure consistency in decision making, and this, in turn, relates to strategic conformance.

From this experiment, we propose that automation confidence interacts with user decision activity—not just in terms of decision outcome but also in terms of information search. However, it is not a simple matter of searching for more information when automation confidence is low. Such a strategy could be counterproductive for the user because it would lead to an increase in workload. Recently, we have reported a model of the optimal use of information sources by healthcare professionals that demonstrates that the strategies for information-search strategy and decision making emerge from the reliability of the information sources, the relative usefulness of these sources, and cost of accessing these sources [2]. From this model and the results reported in this paper, we propose that the ecology of the decision environment (in terms of the availability and access cost of information, and the confidence of support provided to the decision maker) creates a tradeoff space for the analyst. Consequently, rather than searching for more information that the computer might hold but not have used, the user would seek
information from sources outside the computer. The design of a dashboard could, therefore, not only focus on the behavior of the automation (in terms of indicating its recommendation and confidence in that recommendation) but also the availability of information that could be relevant to that recommendation.

As a final point, we note that there continues to be rather limited research into the design and use of dashboards (even though these are growing in popularity). Resnick proposed, in 2003, that there is a need to understand the ways in which dashboards (as high-level summaries of data) should allow users to understand variability in the source data, to ensure that users do not erroneously see patterns in data and should direct users to additional, relevant data [27]. In this paper, we contribute to these aims through the exploration of the ways in which information can relate to different decision tasks, and how user performance can be influenced by “source variability” in the form of reliability of automation recommendations. Further work could explore the decision processes that people make in response to dashboards, particularly when the dashboard seems to be designed to support rapid, intuitive decision making [11], [27].

REFERENCES


