2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications;
Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing

Technology and information trust for supporting risky decisions based on social-media content

Jason R. C. Nurse[†], Michael Goldsmith[†], Sadie Creese[†], Koen Lamberts[§]
[†] Cyber Security Centre, Department of Computer Science, University of Oxford, UK
{*firstname.lastname*}@cs.ox.ac.uk
[§] Department of Psychology, University of York, UK
koen.lamberts@york.ac.uk

***Abstract*—As the availability of open-source information online increases, there are growing concerns regarding its reliability. This has led to renewed emphasis in quality- and trust-metrics research within the social computing space, to assist individuals in determining how reliable pieces of information actually might be. In this article, we take a step back to rigorously investigate the utility of trustworthiness information support provided via computer and information technologies. Our research aim is to assess whether people can cognitively combine trustworthiness advice and evaluative content to make decisions, particularly in a risk-related context. Moreover, we analyse individuals' ability to sensitise their decisions given that information and the criticality of a set task. The results suggest that individuals can perform well at both these tasks even when there are only subtle variations in information and advice. This empirically validated contribution provides a basis for a commonly-made assumption, and reinforces humans as efficient and effective information processors. The study also highlighted several social computing factors that may affect such decisions including quantity of content, existing trust relationships and reasoning behind trustworthiness advice.**

***Index Terms*—Social computing; social media; open-source content; information trustworthiness; decision-making; decision sensitivity**

I. Introduction

Sound decisions are based on reliable information, and that rule also applies in online or virtual environments. Unfortunately, the reliability of online information can be difficult to evaluate, particularly because of the freedom of anyone, anywhere to publish content – typical information contributions can be fact, fiction, opinion or rumour. When online sources of information are used to inform a decision, questions of trustworthiness and quality of the information are critical. Situations that exemplify where trustworthiness information may be useful range from the relatively benign, such as reflecting on product reviews prior to purchase (a task that can be fraught with several issues, as discussed in [1]), to the life-critical, e.g., utilising Twitter content from eye-witnesses (in effect, crowdsourcing) to guide and inform an Emergency Operations Centre's crisis response efforts [2].

To address the problem of unknown information quality and trustworthiness, several quality and trust metrics have been proposed in computer science and social computing [3–5]. Some metrics rely on manually provided information, but others do not require human intervention for their calculation, which is a significant advantage for their effective application in real-time decision-making. Metrics have been based on information features such as recency of content, its completeness or length, content complexity (using readability indices like Flesch-Kincaid), punctuation and typos, grammatical errors, other individuals' feedback on the content's quality, and the authority and reputation of the information's author. Recent research in the social-media field also demonstrates the potential utility of metrics based on the number of unique characters in content, the existence of swear words, pronouns and emoticons, and the number of followers a person has and length of the information author's user names [6]. Clearly, a broad range of information and source features can act as key indicators of the extent to which human users might believe and trust online content.

The value of advice pertaining to the trustworthiness of information content within decision making depends on how people use that advice in their judgement processes. Ultimately, trustworthiness measures or advice must be combined with the evaluative content itself to form robust judgements. In this article, we build on earlier research on the effects of trust on decision making and social interactions (e.g., [7, 8]), and focus specifically on observers' ability to integrate content with trustworthiness advice. Our research contribution is in the investigation of how and under what conditions visually presented trustworthiness advice can improve decision making, especially on risky decisions, based on online content.

Similar research work that has adopted and successfully trialled techniques to convey credibility and quality measures, includes: Idris *et al.* [9] with their traffic light colouring scheme; McGuinness and Leggatt [10] that prompt users with visual alerts; and Volk *et al.* [11] and their trust visualisation based on radar plots and pie charts. Idris *et al.* is particularly noteworthy as our experiments also use this visualisation method with the expectation of benefiting from the real-world traffic light metaphor – use of metaphors to assist understanding is a standard design principle – and innate human perceptual capabilities [12]. The finding that individuals pay attention to the visualisation during decision-making is encouraging as well, as this provides some support for our social computing assessment of the ability to cognitively combine information. What sets our research apart from these and other articles is our assessment of the core human ability to effectively and efficiently make the cognitive combinations

978-1-5090-0154-5/15 $31.00 © 2015 IEEE
DOI 10.1109/CIT/IUCC/DASC/PICOM.2015.66

of these two types of information towards arriving at well-conceived decisions.

In addition to the cognitive assessment, a secondary goal in this paper is to investigate how well people are able to sensitise their decisions given information and associated trustworthiness advice, set in the context of the criticality of a specific task. Depending on the findings, we may be able to apply standard rational and irrational theories, and well-researched heuristics and biases (e.g., [13]) to gain further insight into the decisions made. Generally, if participants are capable of performing both these tasks effectively, this would form a much needed empirical basis for the future use of trustworthiness (and likely quality and credibility) advice within computer decision-support applications. In addition to fulfilling our research project's aims of applying trustworthiness for decision-support, the significance of this research is especially drawn from the increasing number of Web sites attempting to incorporate this and similar advice, typically via feedback statistics or reputational emblems accompanying the information presented to individuals online. Wikipedia and their Article Feedback Tool, seller and reviewer ratings on sites such as Amazon and eBay, and Twitter Verified Accounts are all incarnations of this. Some of these support mechanisms have been researched prior (Wolf and Muhanna [14], for example, assess eBay and Amazon feedback information and how it is interpreted by buyers), but none focus on our specific research aims.

II. The Experiments

A. Research Aims

The first goal of this research is to evaluate humans' ability to cognitively combine information content and the trustworthiness measures that relate to them, to make well-conceived decisions. This evaluation seeks to further validate the preliminary findings of our prior research in [15] with the use of a larger sample set of individuals and a different data set and type of test scenario. The context used for this experiment is intentionally dissimilar to that work to allow us to be relatively confident that our findings are not overly dependent upon context. Another particularly intriguing characteristic of this current experiment is that we now consider the notion of risks and personal safety and decision-making – this therefore moves away from the less personal previous study that considered only product reviews. To allow for this, we draw on the information (Twitter and Facebook posts and news articles) from the UK Riots of 2011 and use this as the foundation of our data set and context decision task. The use of a crisis situation is also beneficial because a prime application of our broader work is supporting situation awareness via computer and information technologies within Crisis Management situations such as riots and disaster zones [16].

A second important goal of our work is to assess individuals' ability to sensitise their decisions based on the criticality of a given task and the information (inclusive of trustworthiness measures) that have been provided in a scenario. This therefore extends the question of, 'can humans combine information and trustworthiness values', to, assuming they can accomplish this cognitive combination, are they able to arrive at well-conceived judgements which also consider how important a related task is? This is a unique assessment which has not been covered in previous social computing work. Formally, these goals lead to two research questions which guide this manuscript's contributions. Firstly, can individuals cognitively combine information content and trustworthiness measures across a number of contexts? Secondly, are persons able to sensitise their decisions based on the criticality of the task at hand and the information that has been provided to them?

B. Participants

43 individuals (21 females, 22 males, Mean$_{age}$ = 28.30, age range: 19–48 years) participated in our study. Participants were recruited with flyers posted within the University of Oxford and University of Warwick, and they included students from a variety of disciplines and levels of study and working professionals, including hospitality clerks, personal assistants, researchers and administrators. Participants were compensated for their participation in the experiment. Initial screening revealed that the participants were experienced in the use of technology, assessing the usage of map-based touch-screen interfaces, and had a relatively normal risk appetite and tolerance. Finally, participants were questioned to determine what portion of them were directly affected by the UK riots in focus – this was done to check for any subsequent overly irregular spikes in scores or opinion. Specifically, only 3 persons noted being affected and their data was still within the normal distribution of scores and decisions.

C. Method and Procedure

The experiments were built around the threat to personal safety faced within a particular geographic scenario, and explored human decision-making based on perceived threat and risk. Thus, given several map-based scenes and geo-tagged information (including Twitter, Facebook and official news posts, and respective trustworthiness measures) describing those scenes, participants were asked to rate each scene on a scale of 1 to 10 – with 1 being the lowest score and 10 being the highest score – based on how threatening or risky they felt the area was.

In addition to providing a single score, participants were also asked a series of questions regarding whether they would travel to the location to shop, to go to work or for a very important medical appointment which took months to arrange; a simple 'Yes' or 'No' to each question was the only response required. Participants were told to view these activities with varying levels of importance such that shopping was to be regarded as a casual activity, going to work was important and more important than shopping, and attending the medical appointment was of utmost importance (i.e., it was the most important activity of the three). We appreciate that depending upon real-world scenarios, the importance of these situations may change (e.g., a doctor may view going to work in a crisis as more important than going to an appointment), but made it clear to participants that they were to use the importance levels provided. These would be used later in the study to assess whether the level of perceived threat and importance of the reason for travel would influence participants' opinion on travelling to the specified location.

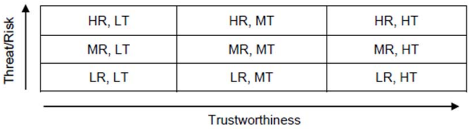
The stimulus materials were presented through purpose-built software on a Motorola Xoom tablet PC. There were several screens presented to participants, each one displaying another scene, which contained a number of pieces of related information. All information content items had trustworthiness measures associated with them indicating to what extent the source or author that composed the information was to be trusted. The task in the experiment was therefore to present this information and ask participants to provide an overall threat rating / score for each screen. Then, they would need to answer the three other travel-related decision questions. Participants were given a maximum of four minutes to read the content on screen and make their decisions.

After completing the rating task, all participants were asked to complete a questionnaire focused on gathering background information and demographic data. Semi-structured interviews were also conducted with a randomly selected subgroup of 20 participants (10 females, 10 males, Mean$_{age}$ = 31.0). This aimed to gather feedback on what motivated participants' ratings and other general motivations and opinions. All experiments were conducted in quiet rooms to avoid interruptions and only involved the participant and experimenter. Experiments lasted for approximately one hour, with participants taking intermittent breaks as they desired.

D. Experiment Design

Identical to our previous work [15], to design the experiment we defined two independent variables: threat / risk level and trustworthiness. Threat / risk level would capture how threatening a piece of information is and could be separated into three levels: Low Risk (LR) – active affirmation that nothing out of the ordinary is happening in a particular location; Medium Risk (MR) – information stating that there is an on-going incident but it is not very serious; and High Risk (HR) – warnings, notices and other information to suggest that there is an increasingly violent situation at a defined location. As such, an example of LR content from our UK Riots data set is, "On Tottenham Court Road, nothing much is happening, it is actually a quite pleasant day". The second variable, i.e., trustworthiness, was used to indicate the extent to which a source of information should be trusted. Again, there were three levels of trustworthiness: High Trustworthiness (HT), Medium Trustworthiness (MT) and Low Trustworthiness (LT). An example use is, "The BBC as an information source has been rated as highly trustworthy". To define the types of content that could be used within the experiment, we then plotted the threat and trustworthiness levels against each other; this resulted in nine possible types – see Figure 1.

For each scene in the experiment, we chose to use two types of content and selected three content items of those types from our riots data set – these two types of content would form a single combination type. This meant that a total of six items of content would be displayed within each scene and content set. This was a reasonable amount of content for the study considering its overall aims, i.e., the research was focused on whether persons can perform the cognitive combination task, not necessarily with emphasis on the volume of information available, but also practical aspects such as the size of the tablet screen and lines of content in each information item (typically three/four lines).



HR, LT	HR, MT	HR, HT
MR, LT	MR, MT	MR, HT
LR, LT	LR, MT	LR, HT

Fig. 1. Matrix plotting the three threat / risk levels (HR, MR, LR) against the three trustworthiness levels (HT, MT, LT). This introduces the nine types of content.

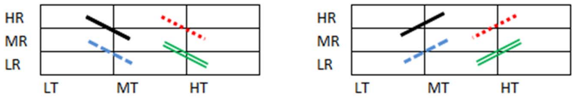
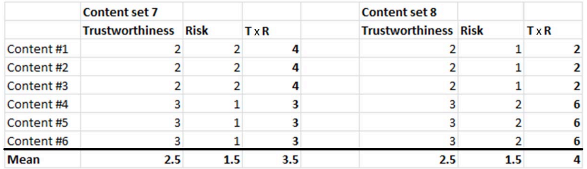


Fig. 2. Each slanted line represents a combination type. The dotted red line in the left side box shows the combination: {{HR, MT}, {MR, HT}}. Colouring and line formatting is used to match subtly different combinations of risk and trustworthiness across types.

There were eight different combination types (and thus, content sets) chosen for presentation to participants, as shown in Figure 2. These combinations were of particular interest because they combined different levels of trustworthiness and threat / risk in a subtly different way, which would lead to different final weighted scores for content sets that were to be compared. Figure 2 displays these combination types and highlights the slope of the matched combination type as the differential factor. In essence, positively sloped combination types (right of the figure) result in higher weighted scores than their negatively sloped (left of the figure) counterparts. Here, we assume and assign a simple ordinal scale to both threat / risk level and trustworthiness where, HR, MR, and LR have scores of 3, 2 and 1 respectively and HT, MT, and LT have scores of 3, 2, and 1 respectively as well. This is an acceptable, albeit arbitrary scale, which allows us to focus our investigation on the correlation effects, towards answering the research questions. Figure 3 displays an example of how information from two combination types is combined (*Trustworthiness* × *Risk*), weighted and compared.

Building on the setup above and the details presented in Figures 1 and 2 therefore, one of the crucial questions that this design allows us to ask is: when comparable content sets (i.e., lines with the same colours in Figure 2) are presented to participants, do they perceive the sets with the positively sloped types as more risky or threatening than those that are negatively sloped? Using the sets in Figure 3 as an example therefore, when participants give their 1–10 threat scores, do they tend to give Content set 8 a higher score than Content



	Content set 7				Content set 8		
	Trustworthiness	Risk	T x R		Trustworthiness	Risk	T x R
Content #1	2	2	4		2	1	2
Content #2	2	2	4		2	1	2
Content #3	2	2	4		2	1	2
Content #4	3	1	3		3	2	6
Content #5	3	1	3		3	2	6
Content #6	3	1	3		3	2	6
Mean	2.5	1.5	3.5		2.5	1.5	4

Fig. 3. Weighted mean calculations for Content set 7 (negatively slopped green double line in Figure 2) and Content set 8 (positively sloped double line also coloured in green).

set 7? This would be the expected outcome as its weighted mean is higher due to greater trustworthiness being placed on higher-risk content. If participants are able to recognise this and correctly assign positively sloped content sets with higher threat scores, then we can be confident that individuals can successfully cognitively combine content and trustworthiness measures towards making a decision, thus fulfilling the first research aim.

The other six content sets (CSs) that formed the basis for the experiment followed the same technique as applied to sets 7 and 8, and therefore their calculations are not presented. For completeness however, and referencing Figure Figure 2, their weighted means are as follows: CS1 (black line and negative slope) produces 3.5 and CS2 (black line and positive slope) produces 4.0; CS3 (red line and negative slope) produces 6 and CS4 (red line and positive slope) produces 6.5; CS5 (blue line and negative slope) produces 2 and CS6 (blue line and positive slope) produces 2.5. To prohibit participants from recognising the underlying design, notion of slopes and thus, attempting to predict scores without properly analysing the content, content sets and the information content items within them were presented in random order.

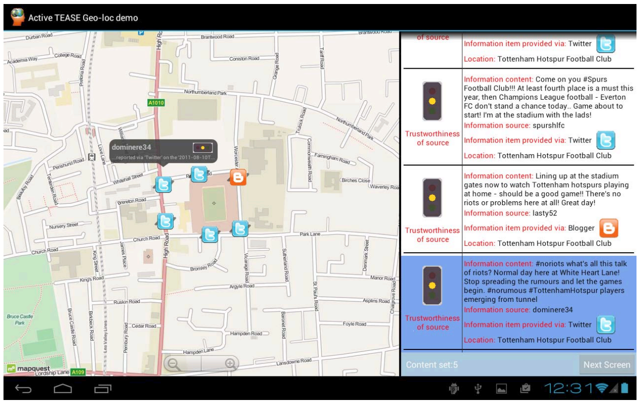
To answer the second research question, the comparison activity introduced above was repeated, but on this occasion, the main consideration was the Yes / No responses from comparable content sets. The specific aspect being assessed therefore was whether participants were sensitive to subtle changes in the actual / perceived risk level such that it influenced their decisions to either shop, work or attend the important medical appointment. For example, suppose for CS3 an individual gave the following responses: Threat level (i.e., the 1–10 score) – 6, for Shopping – 'No', Working – 'No', the Important Medical Appointment – 'Yes', and for CS4 gave: Threat level – 7, for Shopping – 'No', Working – 'No', the Important medical appointment – 'No'. This could indicate that the individual was sensitive to the subtle difference in threat levels of the content sets such that it affected and influenced their resulting travel decision.

E. UK Riot Dataset

To support the experiment, various open-source online content items (e.g., tweets, news agency reports, emergency service data, and so on) from the UK Riots of 2011 were indexed. In total, 31 items that matched the type of data necessary for the threat / risk levels in the respective eight content sets, were selected for use within the experiment. This amount was sufficient given that essentially the same content would be used for content sets that would be later compared; this had the benefit of reducing variation across comparable sets by only changing the trustworthiness of items. Each data item included the type of source application (e.g., Twitter, Blog, Police, Fire Service, or News agency), any username that was provided (most common with Twitter or blogs), the location or area which the information spoke about (e.g., the tweet's geo-tag or the particular street referred to by a news or emergency services report – this metadata was important especially in allowing appropriate item placement on the map), and the content itself (i.e., the information that was published). The content items did not include pictures and for the experiment, participants were told that all of the information related to the same general date and time.

To reduce the likelihood of misinterpretations in the risk levels of content items, a risk level validation process was employed. In this activity, each item was presented to a set of 15 individuals who were tasked with rating its threat level at High, Medium or Low. If more than 85% of people agreed with the assigned threat level, the information item was kept as it was. If less agreed, either a new item of the desired level was selected or the data carefully updated towards the assigned threat level. In these cases, data was re-validated by a different set of 15 individuals until the desired outcome of more than 85% agreement on each item was reached.

F. Experiment Application



of source
Information item provided via: Twitter
Location: Tottenham Hotspur Football Club

Information content: Come on you #Spurs Football Club!!! At least fourth place is a must this year, then Champions League football - Everton FC don't stand a chance today.. Game about to start! I'm at the stadium with the lads!
Information source: spurshlfc
Information item provided via: Twitter
Location: Tottenham Hotspur Football Club
Trustworthiness of source

Information content: Lining up at the stadium gates now to watch Tottenham hotspurs playing at home - should be a good game!! There's no riots or problems here at all! Great day!
Information source: lasty52
Information item provided via: Blogger
Location: Tottenham Hotspur Football Club
Trustworthiness of source

Information content: #noriots what's all this talk of riots? Normal day here at White Heart Lane! Stop spreading the rumours and let the games begin. #norumous #TottenhamHotspur players emerging from tunnel
Information source: dominere34
Information item provided via: Twitter
Location: Tottenham Hotspur Football Club
Trustworthiness of source

Fig. 4. Map-based application used within experiments; this presents CS5. To the left, a map with geo-tagged information items is shown, and to the right, the respective information content and associated trustworthiness levels are portrayed in a list format. Users of the application can tap items on the map and have the related content automatically selected in the list on the right.

Grounded in the experiment design, a map-based tablet application was developed. Figure 4 displays a screenshot of the application and an example of the scenes presented to participants.

For the experiment, traffic lights were used as a visual technique to convey the trustworthiness of each information source. This builds on the findings from our prior research [15] and other work in Idris *et al.* [9]. Hence, red, yellow and green traffic lights are used to portray Low, Medium and High Trustworthiness respectively; the ordering of red, yellow and green also corresponds to the layout of the light visuals, top to bottom. In total, eight screens similar to the one in Figure 4 were presented to participants; these corresponded to the eight content sets. Participants were asked questions on how threatening a scene appeared and their decision to travel based on each of these screens.

III. Results and Discussion

In this section, the results of the experiment are reported and discussed. The structure is such that we first engage in the presentation and analyses of the threat scores given by participants for the scenario and their follow-up travel decisions, then, report on the findings from questionnaires and interviews while also reflecting on any links to quantitative results.

A. Risk / Threat Scores

For the experiment, the three independent variables defined were Threat / Risk level, Trustworthiness and Threat/Trustworthiness correlation, and the dependant variable was the score given to content sets by participants. Instead of focusing on three levels of threat and trustworthiness similar to the experiment design however, for the analysis design we utilised the fact that each content set could be further abstracted to be either high(er) or low(er) in relation to these two factors. These two levels (namely, high or low) would therefore constitute the two possible values for each of the independent variables, hence a 2 × 2 × 2 analysis design. The Threat level/Trustworthiness correlation variable simply refers to the slope of the content set and as defined prior, slopes can either be positive or negative, thus forming the two possible values for that variable. The combination of these variables can also be represented as done in Table I.

Content set	Risk / Threat	Trustworthiness	Slope
1	high	low	negative
2	high	low	positive
3	high	high	negative
4	high	high	positive
5	low	low	negative
6	low	low	positive
7	low	high	negative
8	low	high	positive

TABLE I
Linking the eight content sets to the independent variables and their values.

A repeated-measurement analysis of variance (or ANOVA [17]) was then carried out on the threat scores provided by participants; SPSS was used for our statistical analyses. This ANOVA revealed significant main effects of Threat level, F(1,42) = 143.81, p < .001, MSE = 3.48, Trustworthiness, $F(1,42) = 11.355$, $p < .005$, $MSE = 2.51$, and Threat/Trustworthiness correlation, $F(1,42) = 43.96$, $p < .001$, $MSE = 2.41$. There was also a significant interaction between Threat level and Trustworthiness, $F(1,42) = 50.99$, $p < .001$, $MSE = 1.30$, and between Threat level and Threat/Trustworthiness correlation, $F(1,42) = 16.58$, $p < .001$, $MSE = 1.15$. Here, the conventional ANOVA test statistic is represented by the F value, the p value highlights the statistical significance (i.e., possibility the result is due to chance alone) of the result, and MSE (mean squared error) measures residual variability in results after the treatment effects have been incorporated. None of the other effects or interactions were reported as significant.

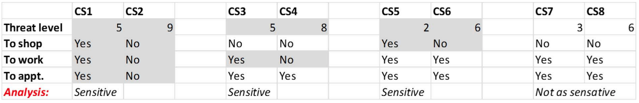
Interpreting the results, there are numerous notable findings which in many ways validate and build on findings in prior work [15]. It was seen that if trustworthiness was lower, participants were slightly less sensitive to the threat / risk level of the content sets. This could indicate that participants regarded this type of content generally as less risky or that they simply ignored some of it. When trustworthiness was higher, the ratings given by participants did reflect the threat level more strongly. Another general finding was that higher trustworthiness content also yielded higher overall ratings than lower trustworthiness (as reflected in the significant main effect of trustworthiness). To compare these findings with those in our previous work however, lower trustworthiness did not have as profound an influence, that is, these information items were not ignored or treated as less significantly in decision-making.

Of most importance for the first aim of this research is the significant main effect of the Threat level/Trustworthiness correlation. If the correlation was negative (i.e., the more risky content items are less trustworthy than the more positive content items), the mean score was lower ($M = 5.03$) than if the correlation was positive (i.e., more threatening content was the most trustworthy, and less risky content items were less trustworthy) ($M = 6.15$). This result is significant as it confirms that participants could cognitively combine evaluative information and trustworthiness advice associated with it in a systematic manner to arrive at ‘expected’ outcomes. This therefore allows a positive answer to be reached regarding the first research question.

Based on a further analysis of the scores, there was no significant male/female difference in choices. That is, neither male nor female was more prone to not being able to identify the subtle differences in comparable content sets. There was also no significant difference as it related to the age of individuals and their performance in the task, or the occupation of participants.

B. Decision Sensitivity

To address the second research aim, a simple analysis of the data received for content sets was conducted. This entailed comparing the threat scores given and travel decisions made by participants when they were presented with subtly different content sets. Where there was a difference in scores (i.e., the perceived threat of a scenario) and a divergence in travel decisions, this was viewed as indicative of the individual being sensitive to subtle changes in the perceived threat / risk level, such that it influenced their decision to either shop, work or attend the medical appointment. Take the data in Figure 5 given by one participant as an example.



	CS1	CS2		CS3	CS4		CS5	CS6		CS7	CS8
Threat level	5	9		5	8		2	6		3	6
To shop	Yes	No		No	No		Yes	No		No	No
To work	Yes	No		Yes	No		Yes	Yes		Yes	Yes
To appt.	Yes	No		Yes	Yes		Yes	Yes		Yes	Yes
Analysis:	*Sensitive*			*Sensitive*			*Sensitive*			*Not as sensative*	

Fig. 5. Threat scores and respective travel decisions given by one participant for the eight content sets presented. Cells with a grey background (viewable in response pairs CS1/CS2, CS3/CS4, and CS5/CS6, above) are used to highlight when the participant was sensitive in their decision. For the CS7/CS8 pair, this gives an example of when the participant was not sensitive (hence no shading).

In comparing the responses to CS1 and CS2, there is a marked difference in perceived threat levels and this is aptly represented in all of the decisions the participant made. CS3 and CS4, and CS5 and CS6 display this progression as well, such that in comparably higher risk situations there was a diminished desire to travel to the affected location. These comparisons and results can be used as a simple indicator that participants were sensitive enough to subtle changes in perceived threat to modify their ultimate decisions. CS7 and CS8 present one example of the contrary case where there was an appreciation of the difference in threat faced within the scenes but their decisions were not changed to suit. This could be because of a lack of sensitivity to these specific scenes and the content within them. Another interpretation could be that the participant regarded the two scenes to be within the same general threat threshold and therefore would react the same way to both. As it is difficult to grasp the specific reason for this (and the amount of CSs to be compared, with the same perceived threat levels), we focus more on the comparisons that suggest sensitivity in decision-making and what was the percentage of these across all study participants.

Comparing the threat scores and travel decisions for subtly different content sets therefore, the results suggested that participants were sensitive. This was apparent in that, in over half (55%) of the comparisons made in total, different perceived threat level scores resulted in different travel decisions. This was true both for situations where perceived threat level scores differed greatly (by more than 4 points) but more interestingly, where they differed slightly (i.e., a difference of 1). Therefore, even when there were small changes, for example, rises in perceived threat level, individuals chose not to travel to the more risky location. This sensitivity was also seen in cases where the importance of the travel decision itself was raised i.e., from travelling to shop to being required to travel for a medial appointment. These findings allow us to conclude that persons do appear able to sensitise their decisions based on the criticality of the task and information (content and trustworthiness measures) that has been provided to them. In essence, a positive results for this research's second research aim.

The data supporting the conclusions above are as follows: there were 172 comparisons (43 participants, each with 4 comparisons), of which 55% indicated a sensitivity in follow-up decisions, 6% were not sensitive, 15% had the same perceived threat level (therefore, decisions were not considered) and 24% showed a difference in threat level but participants had the same decisions. The latter of these points is worthy of further mention because one might regard no difference in decision as an indication of a lack of sensitivity in participants' ability. However, upon a detailed analysis of the data, in these cases threat levels are often very close to each other (e.g., one point apart) or within the same threshold. For example, in many of these cases, we see a CS being given 1/10, the comparable CS being given 2/10, and the resulting decisions for both CSs being Shop – Yes, Work – Yes, Appointment – Yes. It is plausible therefore that participants view these perceived threat scores as so similar that their resulting decisions actually were the same.

Encouraged by the positive findings from the sensitivity analysis, a smaller and more implicit research question regarding sensitivity was then explored, primarily for validation purposes. That is, to assess whether there were any correlations in higher or lower perceived threat / risk and the choice to shop, work or attend the important medical appointment. To investigate this, CSs were first partitioned according to the scores given to them by participants, where 8/10 and above represented a high perceived threat / risk scene and 3/10 and below represented a low perceived threat/risk scene. Next, they were checked for the existence of any correlations in resulting travel decisions. For instance, in scenes rated by participants 8/10 and above, do they mostly only decide to attend important medical appointments and not shop? If so, this could again potentially indicate a link between threat levels, individuals' sensitivity and their final decisions. The diagram in Figure 6 summarises the output of this analysis.

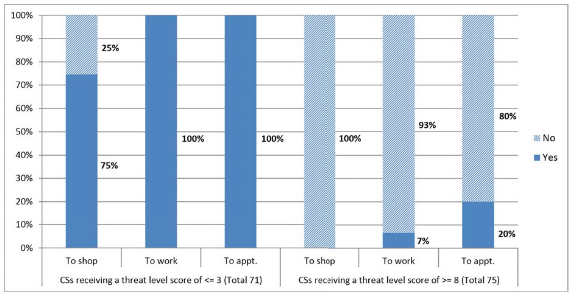


Fig. 6. The differences in travel decisions based on perceived threat scores.

As can be seen in the diagram, the threat scores had a noteworthy influence on travel decisions. This was such that in situations where there was higher threat (level of 8/10 or higher), individuals chose not to shop, only a few chose to attend work but 20% said they would still attend the medical appointment. This again demonstrates the expected influence of threat level on decisions of varying importance / criticality; even when threat is high however, it is apparent that the importance of the appointment does drive participants to still travel there. Where perceived threat levels were low (level of 3/10 or lower), all individuals chose to travel to work and to attend the appointment while only around a quarter chose to shop. This suggests a correlation between degree of risk and resulting decisions made, i.e., there is little risk therefore participants do not hesitate to travel to more critical places such as work and the appointment.

Both of the research results reached in this study are encouraging on a broad scale but especially for the aims of our research project as the scenarios and problem situations used represent real-world decision contexts that the project tool is likely to be deployed in. On the wider scale, results are important because although it is particularly crucial that individuals are able to cognitively combine content and trustworthiness scores (for the experiment, this meant grasping perceived threat based on content and associated trustworthiness), what is also important is their ability to make well-reasoned decisions based on it. In this manuscript's investigation, this capability was tested and largely verified using decision criticality. It was especially encouraging that participants could recognise subtle differences in content sets (which were largely linked to the trustworthiness assignments) and modify their decisions appropriately. The overall correlation in decisions when threat levels were high or low also displayed a good understanding of the content, the task and the significance of judgements based on threat.

C. Interview Findings

To assess the interview data gathered, a simple content analysis and coding technique (as discussed in Berg [18]) was applied to the transcribed interviews. Due to space limitations, we only report on one of the main questions posed to participants, i.e., 'Can you outline and explain your thought process for giving a particular score to a screen (content set)?'

The core aim of this question was to gain an insight into the decision-making process of participants; findings here may act to support or challenge the quantitative findings gathered. In response, the interview results showed that most participants were heavily guided by the trustworthiness measures in making their decisions regarding content and the threat scores they gave. One participant even stated that normally they would not trust Twitter but during the experiment, if a tweet had a green light (indicating HT) associated with it, then they would believe the content fully. There was also indication that a few participants were willing to search for the higher trusted sources first, rather than browsing through all the icons and content on the map in a more natural, sequential fashion. Using the higher trusted content as a basis, they then moved to less trustworthy sources to consider these – although slightly time consuming initially, this technique would have likely enabled better grouping and later comparison of different levels of trustworthy content. This only occurred with a couple of individuals however, as the majority accessed and read through content sequentially as laid out on the map. In both situations, participants resorted to scrolling the side bar after accessing all the icons on the map to remind themselves of the data and trustworthiness levels and to cement their decision. Only then were they comfortable in providing a score to the experimenter.

Another similar non-sequential technique used by participants was to visually scan the map and list for familiar and official sources (e.g., BBC or Oxfam). In one case, a participant noted that he would read that content first and then relate all subsequent messages in the scene to what he had heard from these sources – at times, not even viewing the trustworthiness score assigned by the system to the source. This perspective depicts an expected influence of reputation and prior experiences on personal trust decisions; this is personal because others might have had differing experiences with those sources. This pre-existing knowledge of a source has its advantages but also can be a detrimental at times where official sources are incorrect or outdated (a growing reality as seen in the Mumbai and Boston crises of recent years [19, 20]). Even in situations where a trustworthiness-provision system could pick up on these inadequacies and reduce the trustworthiness level appropriately, a problem that may then arise is, will users trust the trustworthiness score or will they revert to their own opinions and beliefs regarding the source? As such, any such tools in the future will need to consider whether to explicitly offer functionality which supports system users in ignoring preconceptions regarding sources, and in some way draws there attention more to the findings of the metric over their preconceptions through some cognitive persuasion mechanism which could be switched on or off. This could be assessed in further research. To avoid any confusion during the experiment however, we had assigned sources such as the BBC, Oxfam and Guardian as HT.

When asked further about how participants actually came up with their scores on how risky / threatening a scene was, most individuals, particularly those that read content on the map sequentially, said that they read all the content and then decided a score in the end. According to participants, this score was based on their feelings, emotions and initial reactions to the content. This summary view was therefore (apparently) adopted as opposed to consciously calculating a threat level for each message and weighting with its trustworthiness, then combining them as the participant read through the content set. This was an interesting and encouraging finding considering that the quantitative data (even if only localised to the individuals interviewed) does show that this less structured and more emotive sub-conscious technique did still result in expected distinctions in risky and subtly more/less risky situations. This was also with the time constraint placed on the decision-making task. Transitioning from this finding, one avenue for future work would be to determine if this technique still holds for larger amounts of content items that are substantially more diversified in focus and trustworthiness.

As compared to the results from our prior study, there is some similarity in the focus on content from sources rated with higher levels of trustworthiness. What is less clear from this current study's qualitative data, however, is whether there was as predominant an emphasis on higher trustworthy content for scores (i.e., was the decision largely made after reading only that content). In the previous experiment (i.e., [15]), participants could immediately perceive and jump to the higher trustworthy sources in the list but within our current map-based experiment, sequential map access was more natural for participants and that may have affected this type of decision making.

IV. Conclusion and Future Work

The aim of this research was to conduct a social-computing study, to investigate the utility of the quality- and trust-metric values, provided in technology displays, in supporting decision-making. This is especially towards helping persons understand what information to believe and act on when they are making a real-world, risk-based judgement. In particular, we assessed whether individuals can cognitively combine trustworthiness advice and evaluative content to make well-conceived decisions with technology support. From the results gathered, we were able to substantiate and extend the findings from our previous initial work through quantitative and qualitative analyses, and are satisfied that individuals can effectively perform this combination task. This validation was achieved through the use of a different sample of individuals (with more varied vocational backgrounds), a larger participant set (almost three times what was previously used), and a different user task and experimental scenario.

Further to their ability to cognitively combine content and trustworthiness advice, individuals also demonstrated a capability to sensitise their decisions given that (evaluative and trustworthiness) information and the significance of a set travel task. It was especially encouraging that individuals could recognise very subtle differences in content sets (which were largely linked to the trustworthiness assignments) and modify their travel decisions to suit. Traffic lights were viewed by participants as a useful and helpful technique in displaying trustworthiness, which was as expected.

Other interesting findings that can motivate future work included the fact that existing personal trust relationships with sources of content may override the trustworthiness assignments of the system – this was apparent as some individuals mainly focused on known sources and what they knew about them, rather than relying on the trustworthiness indicator. This pre-existing knowledge of a source has its advantages but also can be detrimental on occasions where official sources are incorrect or outdated. In terms of future work therefore, we will need to consider whether to explicitly offer functionality which supports users in ignoring preconceptions regarding sources. This would in some way draw their attention more to the findings of the metric over any preconceptions – through some cognitive persuasion mechanism which could be switched on or off. Allowing toggling is crucial here as a user may well want to use their preconceptions to guide their decisions, whether irrelevant of the impact or as a 'what if' exercise. Either way, the computer system is there to support the user in achieving their tasks and goals.

Acknowledgment

This work was conducted as a part of the TEASE project, a collaboration between the University of Warwick, HW Communications Ltd and Thales UK Research and Technology. The project is supported by the UK Technology Strategy Board's Trusted Services Competition (www.innovateuk.org) and the Research Councils UK Digital Economy Programme (www.rcuk.ac.uk/digitaleconomy).

References

[1] S. M. Mudambi and D. Schuff, "What makes a helpful review? a study of customer reviews on amazon. com," *MIS quarterly*, vol. 34, no. 1, pp. 185–200, 2010.

[2] A. Mills, R. Chen, J. Lee, and H. Raghav Rao, "Web 2.0 emergency applications: How useful can Twitter be for emergency response?" *Journal of Information Privacy and Security*, vol. 5, no. 3, pp. 3–26, 2009.

[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *International Conference on Web Search and Data Mining*. ACM, 2008, pp. 183–194.

[4] K. Peters, Y. Chen, A. M. Kaplan, B. Ognibeni, and K. Pauwels, "Social media metrics – a framework and guidelines for managing social media," *Journal of Interactive Marketing*, vol. 27, no. 4, pp. 281–298, 2013.

[5] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "Tweetcred: A real-time web-based system for assessing credibility of content on twitter," in *6th International Conference on Social Informatics (SocInfo)*, 2014.

[6] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," in *Workshop on Privacy and Security in Online Social Media*. ACM, 2012.

[7] K. Kelton, K. R. Fleischmann, and W. A. Wallace, "Trust in digital information," *American Society for Information Science and Technology Journal*, vol. 59, no. 3, pp. 363–374, 2008.

[8] J. Tang and H. Liu, "Trust in social computing," in *23rd International Conference on World Wide Web*, 2014, pp. 207–208.

[9] N. H. Idris, M. J. Jackson, and R. J. Abrahart, "Colour coded traffic light labeling: A visual quality indicator to communicate credibility in map mash-up applications," in *ICHSST*, 2011.

[10] B. McGuinness and A. Leggatt, "Information trust and distrust in a sensemaking task," in *Command and Control Research and Technology Symposium*, 2006.

[11] F. Volk, S. Hauke, D. Dieth, and M. Muhlhauser, "Communicating and visualising multicriterial trustworthiness under uncertainty," in *12th Annual International Conference on Privacy, Security and Trust*. IEEE, 2014, pp. 391–397.

[12] J. Nielsen, *Designing web usability: The practice of simplicity*. New Riders Publishing, 1999.

[13] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[14] J. R. Wolf and W. A. Muhanna, "Feedback mechanisms, judgment bias, and trust formation in online auctions," *Decision Sciences*, vol. 42, no. 1, pp. 43–68, 2011.

[15] J. R. C. Nurse, S. Creese, M. Goldsmith, and K. Lamberts, "Using information trustworthiness advice in decision making," in *International Workshop on Socio-Technical Aspects in Security and Trust (STAST)*. IEEE, 2012, pp. 35–42.

[16] J. R. C. Nurse, S. Creese, M. Goldsmith, R. Craddock, and G. Jones, "An initial usability evaluation of the secure situation awareness system," in *9th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, 2012.

[17] F. Gravetter and L. Wallnau, *Essentials of statistics for the behavioral sciences*. Cengage Learning, 2010.

[18] B. L. Berg, *Qualitative research methods for the social sciences*, 5th ed. Pearson International Education, 2004.

[19] Guardian News, "BBC admits it made mistakes using Mumbai Twitter coverage," 2008, http://www.theguardian.com/media/pda/2008/dec/05/bbc-twitter.

[20] The Hollywood Reporter, "Boston Marathon Bombing: Rush to Break News Burns CNN, Fox News," 2013, http://www.hollywoodreporter.com/news/cnn-boston-marathon-bombing-mistake-441551.