Using Information Trustworthiness Advice in Decision Making

Jason R. C. Nurse[†§], Sadie Creese[†], Michael Goldsmith[†] and Koen Lamberts[§]
[†] Cyber Security Centre, Department of Computer Science, University of Oxford, UK
[§] Department of Psychology, University of Warwick, UK
{jason.nurse, sadie.creese, michael.goldsmith}@cs.ox.ac.uk, k.lamberts@warwick.ac.uk

***Abstract*—In a society at the brink of information overload, using a measurement of trustworthiness to focus attention and ultimately reduce risks faced by individuals is an increasingly attractive option in supporting well-conceived decisions. As such, this paper seeks to advance discussions on trustworthiness and decision-making research by critically investigating individuals' ability to cognitively combine trustworthiness measures and the information content that they relate to, to make decisions. This is an often assumed reality but one that is lacking focused analysis in the socio-technical field. In our experiments, as we present trustworthiness information using visualisations on a computer screen, we also conduct a secondary assessment of a range of visualisation techniques to determine whether there are any better or generally preferred approaches to support decisions. Findings from both evaluations are relatively positive and insightful, and amongst other aspects, reaffirm humans as optimal assessors and identify a particularly strong dependence on trustworthiness levels in influencing to decision-making.**

***Index Terms*—Information trustworthiness; decision-making; risk communication; trustworthiness visualisation; user studies**

I. Introduction

The amount of information freely available in modern-day society is phenomenal. For instance, a quick search for 'trustworthiness' in Google results in over 12,400,000 hits, substantially more than most individuals would be prepared, or have time to peruse. As professionals and casual users attempt to pick from this glut of informational content to make decisions, a crucial question that they face is what information to trust, and which sources of information should they trust or rely on. This trust problem is exacerbated when assessing Web 2.0 content (e.g., tweets, blogs, posts and wikis), considering that anyone online can be an author, since the customary gatekeepers to publishing, who historically have had some governance over quality, no longer exist in that sphere.

Clearly, consumers of content develop their own strategies for avoiding information overload and high-risk information, usually using some heuristic (consciously or not) for scoping down the sources that they trust [1]. But, there will be some scenarios in which relying on known or habitual sources will not suffice. For example, crisis situations where time is critical and intelligence very limited, or even simpler scenarios where one is trying to navigate a plethora of hyped marketing and highly mixed reviews about a soon-to-be-released technology gadget. In such circumstances, decision-support which seeks to convey the potential trustworthiness of information will be a crucial factor in avoiding potential negative consequences should the wrong (i.e., high-risk) information be relied upon. Existing approaches to deal with this problem rely on trusted third-parties who investigate the trustworthiness of Internet sources and create a trusted network of content publishers for customers to access. This has a significant limitation in that you are restricted to preselected sources, and so it would not scale appropriately to the scenarios we are considering.

The TEASE research project is aimed at addressing this capability gap by providing information-trustworthiness decision-support for open sources in general, using software which communicates via the visual interface (whether smartphone, tablet or desktop). Through the notion of trustworthiness – its measurement also a core focus of ongoing research work commenced in [2] – we are also able to implicitly communicate the risk associated with using this open-source data. One of the research questions we are addressing is how best to communicate this to a user, as we do not want to introduce cognitive load which could negatively impact the decision-making process, given that TEASE is primarily about helping users to enhance the quality of decisions. We present here our first experiment aimed at testing our research hypothesis that it is possible for a human to cognitively combine trustworthiness measures (of the type defined in TEASE) and the information content that they relate to and make well-conceived decisions. A secondary aim of this paper is to investigate a number of visual schemes for presenting trustworthiness to individuals. The objective in this case is to discover whether there are better or preferred techniques, which can then be widely applied.

The reminder of this paper is as follows. In Section II we review the related work in the field of trust, decision-making and visualising trustworthiness. Section III presents and details the experiments conducted to investigate the goals as mentioned above. Next, we report on and discuss the results and findings in Section IV. The paper then concludes in Section V with key points and directions for future work.

II. Related Work

The significance of trust within decision-making has been studied in various contexts, including information science [3], for purposes of e-commerce and Web-based health advice [4], and within organisations and business [5]. There has also been noteworthy research conducted in the military domain assessing the influence of trust and distrust in decision/sense-making as it pertains to intelligence compilation tasks [1, 6].

These articles provide a glimpse into a vast field of research.

Considering its importance, several researchers, also in the socio-technical field, have proposed models for trust which aim to characterise it and give insight into key factors and influences. Pickard et al. [7] provides one example of such work that assesses the Web environment and proposes a model for trust and understanding trust decisions there. This model contains a number of the generally accepted factors affecting trust but categorised differently (according to internal, external and user's cognitive state) and occasionally specialised for the Web environment (e.g., trust seals and Web site certifications). In [8], Gil and Artz introduce the notion of content/information trust and define various factors which influence it. Though an analysis of these factors, they also highlight a few which may be the most important to individual's trust and decision-making processes. Other pertinent works which cover these topics in more detail include [2, 9].

Beyond factors and models, there have been attempts to quantify trust and trustworthiness. Further to their work on a trust-perception model, Costante et al. [10] highlight the possibility of future extensions which might allow the measurement of trust factors and ultimately, the quantification of the trustworthiness of Web sites. Within the social-media domain, Moturu and Liu [11] present positive results to support their approach to quantify content's trustworthiness. Their proposal is based on an unsupervised, feature-driven technique and is composed of numerous scoring models for quantification. According to [11], the key aspect of their proposal is its ability to be applied across various social media. These works provide insight that will influence our other ongoing research towards defining more comprehensive trustworthiness measures.

Assuming a situation where reliable quantification is achievable, the next natural step is presenting this trustworthiness level to users to facilitate better decision-making regarding the use of information. An example of research that has attempted a goal akin to this is [12]. There, the authors conduct a study on presenting an information source's credibility rating using a traffic light visual indicator to generally assess how this may influence users. One of the more relevant outcomes of their work is that such visuals were viewed by participants as important in influencing their credibility-based decisions.

Visually presenting trustworthiness levels is also a goal of our research. Within the literature, there are some proposals towards this goal (such as [13,14]) but further research is needed in critical evaluation of proposals. The field of uncertainty visualisation might be a useful place to draw inspiration as this is supported by several years of research. Pang et al. [15] posit that the ultimate goal of this field is to provide individuals with visualisations that reflect uncertainty (errors, variations, noisy or missing data) information to assist in analysis and decision-making. In that article [15], a comprehensive survey of uncertainty visualisation techniques is also presented. These span from the simple approaches utilising colour, shapes and blinking, to the more complex glyphs, contour lines and animation. More recent studies [16, 17] have concentrated on evaluating some of these and other visualisation approaches, at times with the aim of supplying general guidance for designers to use. Bisantz et al. [16] in particular, identify transparency, brightness and saturation as useful techniques to convey uncertainty provided that background images and the overall task context are considered.

III. The Experiment

A. Research aims

In line with the discussions above, our experiment has two aims. The first and core aim is to validate individuals' ability to cognitively combine trustworthiness measures and the information content that they relate to, to make well-conceived decisions. The second aim is more specific than the first and seeks to investigate visual schemes for presenting trustworthiness. The goal in this case is to discover whether there are better or generally preferred techniques for assisting people in decision-making.

B. Method and Procedure

The experiment was framed in the format of a simple product rating scenario. Participants were asked to rate a particular product (the Samsung Galaxy Note) on a scale of 1-10 (with 1 being the lowest score and 10 being the highest score) based only on the information given about it on a screen and in a limited time (5 minutes). The information on each screen consisted of six third-party reviews of the product, each with an associated trustworthiness measure visually depicted. The screens drew from a set of real but controlled product reviews devised specially for this experiment. The trustworthiness measure associated with each review indicated to what extent the reviewer who wrote the review was to be relied upon and trusted by the participant. The first part of the experiment (Part A) focused on simply presenting these two aspects and asking participants for an overall rating/score for the product. The second part (Part B) varied the ways in which trustworthiness measures were presented, and again, participants were asked to provide only an overall rating after viewing each screen. Each part had a total of eight screens.

After completing the hands-on task, an interview was conducted with participants (and audio-recorded). The interview gathered feedback on what motivated the ratings and subjective opinions on the effectiveness of the visualisations. All experiments were conducted in isolated rooms to avoid interruptions and the hands-on task entailed participants using a 10" Motorola Xoom tablet PC. Experiments lasted for roughly 1 to 1.5 hours with breaks.

Fifteen individuals took part in the experiment. They were comprised of a mixture of postdocs, and postgraduate and undergraduate university students. The average age was 24.2 years (a Standard deviation of 5.24). Seven males and eight females participated, and they were from a variety of disciplines spanning Social Sciences, Arts, Medicine and Science.

C. Design

1) Part A (Research Aim 1): The design of the experiment involved two independent variables, namely trustworthiness

and positivity. We defined three levels of reviewer trustworthiness: High Trustworthiness (HT), Medium Trustworthiness (MT) and Low Trustworthiness (LT). An example use is, "A reviewer has been rated as highly trustworthy". Also, there were three levels of review (information) content positivity: Positive (P), Neutral/Okay (M) and Negative (N); for example, "Product X was a horrible purchase! It was very overpriced and did not live up to expectations at all!", i.e., a Negative review. Combining trustworthiness and positivity yielded a matrix of nine types of review (determined by content and trustworthiness), as shown in Figure 1.

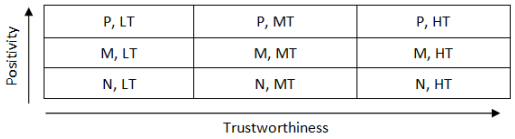

Fig. 1. Trustworthiness and positivity matrix

During each product scoring task, only two cells were populated, with three reviews in each cell, ensuring that a total of six reviews would be displayed on screen. This arrangement therefore defined exactly what level of positivity and trustworthiness was necessary for each review displayed. Six reviews was a manageable number considering that review information was a few lines long, that we were using a tablet PC for experiments (therefore a smaller screen), and finally, that there were several screens shown to the same users.

In total, eight review combinations were chosen, with each one spreading two cells as discussed above; an example of a combination is {{N, LT}, {M, MT}}. Particularly intriguing areas which we intended to test (i.e., the motivation for the combinations selected) were where combinations were adjacent and merged two different levels of trustworthiness and positivity in a subtly different way, which led to different actual weighted score calculations; this is discussed further below. It was therefore worthwhile to assess whether individuals picked up on these differences. In terms of visuals, the traffic light colour spectrum was selected to present the trustworthiness measures across all eight screens. Thus, green, amber and red represented HT, MT and LT respectively for the reviewer. In Figure 2, the review set combinations chosen are portrayed and colours are used to match the combinations that possess subtle differences and are therefore of most interest.

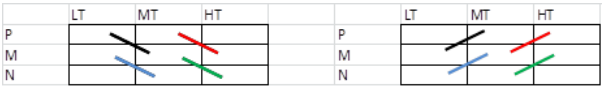

Fig. 2. Combinations chosen to provide Part A experiment data

It is worth pointing out that on the left of Figure 2 there is a negative slope/slant matrix, and on the right there is a positive slope/slant matrix. Positive slopes result in higher weighted scores as will be shown, and therefore should result in higher user scores when compared to their respective negative slopes (matched by colour). In order to allow us to investigate correlation effects, we assign an arbitrary ordinal scale to both trustworthiness and positivity. This is as follows. For Trustworthiness, Low Trustworthiness (LT) = score of 1; Medium Trustworthiness (MT) = score of 2; High Trustworthiness (HT) = score of 3. Whilst for Positivity, Negative (N) = score of 1; Neutral/Okay (M) = score of 2; Positive (P) = score of 3.

To consider, as an example, the two red lines from Figure 2 (which signify (i) on the left, {P, MT} and {M, HT} and (ii) on the right, {M, MT} and {P, HT}) and the numeric values for trustworthiness and positivity above, the following weighted means in Figure 3 could be reached. (This combination could have been made in various ways, the main purpose is to show where we should expect a difference in scores.) Considering these final values, when presented to users, the screen with Review set 4's data (with a value of 6.5) should therefore result in a higher user score than Review set 3's data (which has a value of 6). Hence, the question is, can this actually be seen in the experiment data provided by users during tests? It should also be noted that we randomised the order of the reviews on screen and the sequence in which screens themselves were presented. This stopped participants from being able to approach the task in a systematic fashion and predict scores without properly analysing the data.

	Review set 3				Review set 4		
	Trustworthiness	Positivity			Trustworthiness	Positivity	
Review #1	2	3	6		2	2	4
Review #2	2	3	6		2	2	4
Review #3	2	3	6		2	2	4
Review #4	3	2	6		3	3	9
Review #5	3	2	6		3	3	9
Review #6	3	2	6		3	3	9
Mean	2.5	2.5	6		2.5	2.5	6.5

Fig. 3. Calculating weighted means in Review sets 3 and 4

As the other review sets possess similar calculations to deduce the weighted means, these calculations are not presented here due to limitations in space. In summary, these means are as follows: Review set 1 (black line, negative slope) is 3.5 and Review set 2 (black line, positive slope) is 4; Review set 5 (blue line, negative slope) is 2 and Review set 6 (blue line, positive slope) is 2.5; Review set 7 (green line, negative slope) is 3.5 and Review set 8 (green line, positive slope) is 4. As necessary, one can reflect on Figure 2 to determine the trustworthiness and positivity levels for each review set.

2) Part B (Research Aim 2): The design of the Part B was the same as Part A with two main differences. Firstly, there was a variation in trustworthiness visualisation methods, and secondly, different review set combinations were selected. We now present these differences in detail.

To assist in the fulfilment of the second research aim, we initially considered numerous ways in which trustworthiness could be visually conveyed on screen. We settled on four techniques which we believed covered a wide spread of types, which we could then evaluate; most of these have also been used before in the very related field of visualising uncertainty [15, 16]. These were: (i) a traffic light colour spectrum with green, amber and red indicating high, medium

and low trustworthiness of the reviewer respectively ([18]); (ii) a test-tube which was filled higher to represent greater trustworthiness, lower for less trustworthiness; (iii) a star shape which grew with higher trustworthiness and shrunk with less trustworthiness; and (iv) greying out the review itself, where, HT was normal, MT was 30% greyed out and LT was 70% greyed out; akin to transparency use in [16]. Thus, the question of interest for our research was: what was the best way of presenting trustworthiness information to enable individuals to understand it as easily as possible and make decisions? Furthermore, is there a general preference in methods?

In detail, two combinations of trustworthiness and positivity (i.e., review sets) were chosen in addition to the four different trustworthiness presentation techniques. This led to eight screens of reviews. As with Part A, the focus was on adjacent combinations which led to subtle differences, slopes/slants had the same meanings, and there was randomisation in the screen/trustworthiness presentations and the reviews. Figure 4 displays the selected combinations of review types.

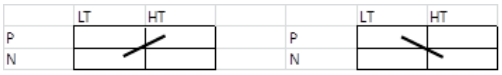

Fig. 4. Combinations chosen to provide Part B experiment data

As we were only using two review sets, there was repetition in basic data sets but differences in presentation techniques. Figure 5 presents the two sets and the resulting weighted means. For this part of the experiment, the objective was to compare these adjacent sets to determine which gave the largest difference in scores. This difference was an indicator (supplemented by interview findings) as to which presentation technique was the most useful at enabling individuals to cognitively process and feedback on the information that had been shown to them.

	Review set 1	Using "Traffic lights"		Review set 2	Using "Traffic lights"	
	Trustworthiness	Positivity		Trustworthiness	Positivity	
Review #1	1	1	1	1	3	3
Review #2	1	1	1	1	3	3
Review #3	1	1	1	1	3	3
Review #4	3	3	9	3	1	3
Review #5	3	3	9	3	1	3
Review #6	3	3	9	3	1	3
Mean	2	2	5	2	2	3

Fig. 5. Calculating weighted means in Part B review sets

Further to Figure 5 which pertains to traffic lights, the other sets and means are as follows: Review set 3 (using greying out, positive slope) is 5 and Review set 4 (using greying out, negative slope) is 3; Review set 5 (using a test tube, positive slope) is 5 and Review set 6 (using a test tube, negative slope) is 3; Review set 7 (using a star, positive slope) is 5 and Review set 8 (using a star, negative slope) is 3.

D. Prototype screenshot

In line with the experiment design above, we implemented a prototype application which could be run on an Android tablet PC. Below are screenshots of the app. To depict the reviewer's trustworthiness we utilised traffic lights and greying out.

Fig. 6. Using traffic lights and greying out to convey trustworthiness

IV. Results and Discussion

In this section we report on and discuss the results and outcomes from the experiments conducted. In the first instance, this surrounds the quantitative analysis, which is largely based on user scores given by study participants for the product. Next, the emphasis shifts to the more qualitative work, which primarily embodies the data from interviews. Where applicable, we reflect on the potential links between these areas.

A. Product scores

1) Part A: After gathering the scores from the experiments, these were input into a table and then fed into the SPSS statistical package. We then carried out a repeated-measurement ANOVA (analysis of variance) [19] on the scores provided by the participants. The independent variables were Positivity (Low or High), Trustworthiness (Low or High) and slope, or more accurately, Positivity/Trustworthiness correlation (positive or negative). The ANOVA revealed significant main effects of Positivity, $F(1, 14) = 39.16$, $p < .001$, $MSE = 2.09$, Trustworthiness, $F(1, 14) = 10.854$, $p < .005$, $MSE = 0.94$, and Positivity/Trustworthiness correlation, $F(1, 14) = 52.60$, $p < .001$, $MSE = 1.31$. There was also a significant interaction between Positivity and Trustworthiness, $F(1, 14) = 14.56$, $p < .002$, $MSE = 0.783$. Here, the F value is the conventional ANOVA test statistic, the p value expresses the statistical significance of the result (i.e., the likelihood that such a result is due to chance alone), and MSE (mean squared error) measures remaining variability in results after the treatment effects have been taken into account. None of the other effects were significant.

A point worthy of mention here is that although this analysis design was based on the work from Section III-C, instead of focusing on three levels of positivity and trustworthiness, we utilised the fact that each review set could be abstracted further to be either high or low in relation to these two factors. These

two levels (high or low) were therefore used to constitute the two possible values for each of those independent variables. The Positivity/Trustworthiness correlation simply refers to the slope of the review set; readers should refer to Section III-C for explanation of slopes.

Interpreting the results above, there were several aspects worthy of note. Firstly, based on the data, it was seen that if trustworthiness was lower, participants were not sensitive to the overall positivity of the reviews. This suggested that individuals saw these reviews generally as less positive or that they simply ignored the review content, both of which were interesting findings that should influence any future design work or experiments. McGuinness and Leggatt [1] provide some comparable feedback from their study participants, where it was noted that when individuals perceived information as 'probably unreliable' (analogous to our low trustworthiness scale), it was immediately set aside or provisionally rejected. In cases where trustworthiness was higher, the participants' ratings in our work did reflect the positivity of the reviews. This difference is depicted in Figure 7 where the Low Trustworthiness (Low TW) line is flat across positivity scores whereas the High Trustworthiness (High TW) line is positively sloped. Another general finding was that High TW also yielded higher overall ratings than Low TW (as reflected in the significant main effect of Trustworthiness). In the former situations therefore, people generally rated the product as better notwithstanding the positivity score.

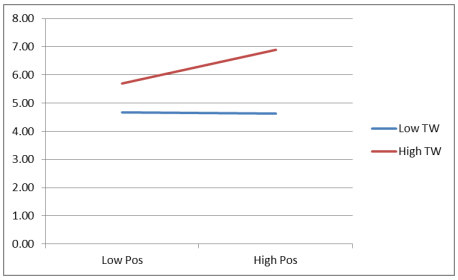

Fig. 7. Means underlying Trustworthiness and Positivity interaction

Most important for our purposes was the significant main effect of the Positivity/Trustworthiness correlation. If the correlation (or slope as referred to in Section III-C) was negative (i.e., the more positive reviews were less trustworthy than the more negative reviews), the mean rating was lower (M = 4.72) than if the correlation was positive (i.e., positive reviews were the most trustworthy, and less positive reviews were less trustworthy) (M = 6.23). This crucial result showed that the participants could combine trustworthiness information with evaluative information in a systematic manner, as one would expect from optimal assessors. Reflecting on our experiment aims in Section III-A therefore, this resulted in a positive finding for the first aim.

2) Part B: The second experiment used a 2 x 4 repeated measures design, with Positivity/Trustworthiness correlation (positive or negative, i.e., 2 levels) and Presentation Mode (4 levels) as independent variables. A repeated-measurements ANOVA on the mean ratings yielded only a significant main effect of correlation, $F(1, 14) = 110.84$, $p < .001$, $MSE = 4.55$, with much higher mean ratings for positive correlations (M = 8.0) than for negative correlations (M = 3.9). This confirms the effect of correlation that was observed in Part A. There were however no significant differences in the effectiveness of the four presentation techniques. Placing this finding into context therefore, that is, as it relates to this research's second aim, there was no 'best way' of presenting trustworthiness according to the scores data that was gathered.

B. Interviews

This section presents and discusses the findings from the interviews conducted. To structure our discussion, we used the interview questions and therefore adopted a question-by-question analysis of findings. In terms of data collection process and analysis methodology, a content analysis technique was applied to the recorded, then transcribed interviews. A semi-structured interview process was preferred to provide a general format, while also allowing for flexibility in exploring any interesting avenues that arose during discussions.

Question 1: What does trustworthiness mean to you?

In assessing the responses to this question, it was apparent that all of the study participants had a clear grasp on trustworthiness and what it meant. In general, the most common description encompassed the notion of reliability, and that the information or person could be relied on to provide good, unbiased information and that it would do what it said it would do. Some of the other, more prevalent comments also highlighted several of the trustworthiness factors identified in previous work in [2, 8]. For example, participants said that something or someone (generally a source) is trustworthy where: there is evidence to what is being said, the source is an authority on the matter and that they are competent, there is a positive motivation to help and not deceive (no hidden agenda), there is no bias, accuracy and truth to what is being said, and that it can be believed and accepted without criticism.

Another point that emerged from the findings was an explicit link, by about half of the participants, to decision-making. Within their definitions of trustworthiness, they noted that if something or someone was trustworthy, they would act on what was said. The content of the information or the message from the individual would therefore form a key component their decision-making process. One participant stated that they would take the information seriously and plan their actions based on it. If the source was not trustworthy, another individual expressed that they would need to be much more wary about what was being said, whereas another participant said that they would not rely on it but would seek out other information to confirm or refute it. Generally therefore, this acted to confirm existing research on the direct link between trustworthiness and decision-making.

Question 2: Can you outline and explain your thought process for giving a particular score to a screen of reviews?

A majority of participants adopted an approach that consisted of the following activities: (i) first, scanning the screen for the reviews by the HT reviewers; (ii) reading these reviews and deciding a score to give the product; and finally, (iii) possibly considering the reviews by the LT reviewers and if they agreed, possibly modifying the score upwards or downwards to reflect their opinion. This thought process was broadly applied across all four presentation techniques. This was a useful finding for our purposes as it showed that generally individuals tended to focus on the HT information first and that this information largely formed the basis of their decisions. Even though LT information might be read afterwards, the decision was mostly already made. For completeness, we note that only two of the participants decided to go through the review information sequentially (i.e., from top to bottom) but they did not give any reason for this preference. This might simply therefore be their preferred reading style.

Some participants also said that they paid special emphasis to the quantity of the types of review to assist their decision, i.e., the number of HT reviews versus MT or LT reviews. Furthermore, reviews of the same trustworthiness type were checked to see if they agreed before coming up with a score for that trustworthiness set. If there was agreement about the product within the same set, this would give more strength to the score. For example, if all the reviews associated with HT were very positive, a score of 9 might be given, however, if in that HT set, only one was positive and the (two) others were neutral, then a score of 7 or 8 might be provided.

Question 3: Part B of the experiment considered several methods of presenting trustworthiness measures. Did you have a preference in methods? Why or why not?

All but one of the participants expressed a preference in trustworthiness presentation methods. This preference is summarised in Figure 8.

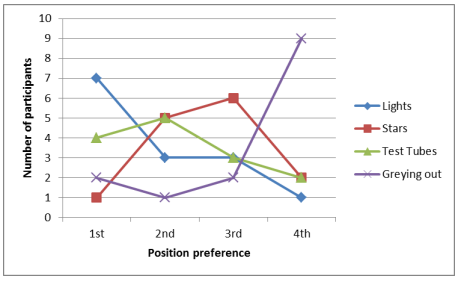

Fig. 8. Summary of the methods of preference

From this figure, we can see that traffic lights were selected as 1st by the most number of participants; specifically, half of the test group. Conversely, greying out was least preferred by participants with 9 of them rating it 4th in their preference lists. Although not directly supported by significant quantitative findings in Section IV-A2, this did give us some idea on preference of methods. Reflecting on the data, the reasons why traffic lights were said to be preferred link to instinctive behaviour (these lights have been central to society for drivers and pedestrians alike for centuries) and colour difference (green and red colours were viewed as striking and very clear). An individual also stated that it was ingrained in people's minds that green is good, but more so, that red is bad. As it relates to the literature, this finding provides some evidence to support the often made assumption (e.g., [12,18]) that traffic lights are a good technique to convey trust and credibility.

For the greying out technique, individuals reported being very unsure about the depth of greying out and what it meant, i.e., whether it was MT or LT. The problem here, therefore, was translating a greyed out screen to its respective trustworthiness measure to understand how bad the trustworthiness really was. Comparing this finding to existing related research, [16] actually found the opposite when assessing the effectiveness of transparency (not that different from greying out) in communicating uncertainty on maps. One reason for this might have been that, unlike our study, all levels of transparency were displayed on screen at a time therefore individuals could assess these and rank accordingly. Returning to our study, in situations where the greyed out method was preferred, individuals stated that they liked the fact that they could very quickly associate bright content with HT reviewers, these reviews also stood out much more. Moreover, there was no need to check a separate trustworthiness graphic, read the review and then combine them to arrive at a score for each review. This task was somewhat implicit as the trustworthiness was directly influencing the way the review was presented.

According to participants, test tubes were also a relatively good presentation method. One of the main reasons for this was that it was clear: a full test tube was HT, a medium one, MT, and an almost empty one, LT. Another reason was that it could allow more precision than the other methods; one could for example imagine representing scores of 1-10 with the test tube but not with the traffic light. The potential issue with the test tubes as raised by roughly half the participants was that they seemed out of context and inappropriate for general scenarios which were not of a scientific nature. This is a salient point and worth consideration in future designs.

Stars tested the idea of growing and shrinking shapes to represent trustworthiness. These were received less well by individuals as compared to lights and test tubes but reasonably okay in a general sense. Drawing from the data gathered, the benefit of stars was that one did not need to remember the meaning of aspects such as colours (e.g., with traffic lights) as it was immediately clear that big stars are HT and little stars are LT. Stars are also a very common rating method today, and therefore, use of stars already had a firm basis. The reason given as to why stars did not feature higher in the preference listings was simply because of a preference to other methods.

Reflecting on all four methods, there were a few interesting findings related to the use of LT representation methods and their influence on participants and their scores. In general, participants were seen to react quite negatively to the LT reviewers, even at times disregarding their content. This was found in common in both parts of the experiment. This may therefore provide an answer to the quantitative findings as it relates to difficulty grasping positivity in LT reviews. An

interesting aspect regarding the use of greying out in particular was that a few individuals expressed that their minds wanted to automatically ignore reviews from LT reviewers because it was difficult to see and read. This instinctive response was not always welcomed, as one participant said that she felt that it was robbing her of what someone less trustworthy had to say. In cases where time was short, she admitted that it would be preferable as it assisted in quickly filtering LT reviewers. However, in the general case, she desired to not be subconsciously influenced, but rather to be able to read the reviews completely, then apply the trustworthiness measures and make her own decision.

Also pertaining to content from LT reviewers, we found that for one participant he saw the low test tube and thought to disregard the review entirely. Conversely, when looking a red traffic light, he thought it might be 'not that bad' and therefore read the review considering it was 1/3. The difference was that the low test tube seemed like it was almost empty (around 5%), whereas a red traffic light could be anywhere between 1% and 33%. This raises a question regarding presentation methods and what level of trustworthiness detail should be presented to users; we expect that context will also play a part because one would presume that certain scenarios need a more detail given what is at stake (e.g., in a crisis management situation), but for more relaxed situations (e.g, product review scenarios), utilising three levels might be sufficient - this is discussed further in interview question #4 below.

Question 4: What is your opinion on the usage of three abstract categories to represent trustworthiness as opposed to more detailed methods, e.g., a percentage from 1-100%?

Slightly over half of participants expressed a desire for more detailed techniques for presenting trustworthiness. They felt that three levels was a little too restrictive and would not convey enough information on trustworthiness for them to make a proper decision. Rather than opting for a percentage score – which they felt would be too much detail – a few of those individuals did mention that five levels might give the best balance between abstraction, simplicity and detail. This therefore led to the test tubes being noted by them as the one technique that may be most appropriate, given that with the others it may be increasingly hard to tell the difference in levels; the traffic lights would not be usable at all in this case.

In cases were the participants agreed with the three-level structure currently used, the reason they gave was because of its simplicity. Having to only deal with three trustworthiness levels was also thought to be useful when faced with a great deal of information. Three levels would be likely to speed up the time spent thinking as less detail needed to be considered. Lastly, there were some participants who appreciated both detail and abstraction and stated that the levels depended on context. Thus, when a serious decision needed to be made it would be better to have the detail, while in other cases, the basic three-level system was preferred as it was easier to read and quicker to assimilate. This was noted as a sensible way to proceed, or even allowing users of a system which implemented these techniques to be able to choose which approach they preferred and when they preferred it.

Having assessed the findings from the complete set of interview questions, we reflected briefly on the preference of trustworthiness methods held by individuals. Our aim was to compare the preferences stated by individuals (gathered in interview question #3) to the actual data scores which they gave each review set (in Section IV-A2), to determine whether they were similar. The product scores provided a useful data outlet because they implicitly captured which trustworthiness methods allowed for better scoring by the individual, i.e., which allowed lower review sets as defined in Section III-C2 to be seen as lower (therefore deserve less scores) and higher review sets as higher (thus deserve higher scores).

	Stated as 1st	Facilitated Best Scoring	Agreed?	Stated as 4th	Lead to Worst Scoring	Agreed?
Participant	1 Test tubes	Test tubes	✓	Greying out	Greying out	✓
	2 Traffic lights	Traffic lights	✓	Greying out	Greying out	✓
	3 Indifferent	Greying out/Stars	n/a	Indifferent	Traffic lights/Test tubes	n/a
	4 Traffic lights	Indifferent	n/a	Greying out	Indifferent	n/a
	5 Greying out	Traffic lights/Test tubes	×	Test tubes	Greying out	×
	6 Traffic lights	Traffic lights/Test tubes/Greying out	✓	Greying out	Stars	×
	7 Traffic lights	Traffic lights	✓	Greying out	Greying out/Stars	✓
	8 Traffic lights	Traffic lights/Test tubes/Stars	✓	Greying out	Greying out	✓
	9 Stars	Stars	✓	Traffic lights	Test tubes	×
	10 Traffic lights	Indifferent	n/a	Stars	Indifferent	n/a
	11 Greying out	Greying out	✓	Test tubes	Traffic lights/Test tubes	✓
	12 Test tubes	Test tubes	✓	Stars	Traffic lights/Stars/Greying out	✓
	13 Test tubes	Indifferent	n/a	Greying out	Indifferent	n/a
	14 Traffic lights	Greying out	×	Greying out	Traffic lights	×
	15 Test tubes	Traffic lights/Stars	×	Greying out	Greying out/Test tubes	✓

Fig. 9. Comparing actual data scores with stated preferences

Combining the actual scores data collected with the stated preferences from interview question #3, we produced the comparisons in Figure 9. From this we observed that in a more situations than not, the stated preferences and actual scores agreed. Participants could therefore generally identify which trustworthiness presentation techniques worked best for them in making the required decisions. Combined with findings above therefore, this did act to reaffirm some level of preference in techniques.

V. Conclusion and Future Work

In this paper, we investigated individuals' ability to cognitively combine trustworthiness measures (that were presented visually) and the information content that they related to, to make well-conceived decisions. We also assessed a number of trustworthiness visualisation techniques with the aim of determining whether there were any better or generally preferred methods. Our general findings were:

- Individuals can combine trustworthiness information with evaluative information in a systematic manner. This confirms existing assumptions but now provides some much needed experimental evidence. In the future, we will seek to build on these findings using a differently framed experiment (likely within crisis management), a more diverse set of participants and a greater sample size.
- When trustworthiness is low(er), individuals do not appear to be sensitive to the overall positivity of the information content displayed. One explanation for this is that they largely ignore information when it has this association. In situations of higher trustworthiness, persons appeared to rate products better overall notwithstanding

positivity. Both of these findings demonstrate quite a profound effect of perceived trustworthiness on decisions.
- Information associated with highly trustworthy sources is normally read first and decisions formed largely based on it. Individuals then, at times, tend to read information from lower trustworthy sources and make slight changes. Moreover, the quantity of types of review (i.e., whether HT, MT, or LT) and whether those types are internally consistent is also a factor when making decisions.
- Out of the four techniques tested, traffic lights tend to be most preferred as a way to represent trustworthiness, whereas greying out is generally least preferred; this was not apparent from the quantitative analysis but rather the qualitative work. The main issue with greying out appeared to be the difficulty in understanding/seeing what degree of trustworthiness was being presented. A prime use of greying out nonetheless does still seem to be in (subconsciously) directing users to/away from content according to its trustworthiness level.
- There is a slight preference for more detailed methods of presenting trustworthiness, particularly where important and critical decisions rely on knowing those details. A 1-5 level seems to be preferred by some study participants, therefore we may experiment with such an approach.

Finally, in addition to the above, there are various questions which future analyses will aim to address; these include:

- The fact that people appear so willing to rely on lower trustworthiness values as a reason to disregard information content is a very strong result for work in this field, as it means that interface designers can certainly help people focus away from what them (or a system) determines is a LT source or piece of information. This does mean, however, that designers of systems have to take responsibility for ensuring that a LT score in particular is shown only at apt times. A set of interesting questions which arises from this discussion is: what happens when we present people with all three trustworthiness levels on screen at once? Is the implicit preference for higher trustworthiness reviewers maintained? Based on current findings, we might expect LT to be completely disregarded, but what weight is placed on MT reviews in decision-making then?
- In one of the summary points above, the quantity of types of review (i.e., whether HT, MT, or LT) and whether those types were internally consistent was identified as a factor when making decisions. Following on from this, another set of intriguing questions for future work are: Could a high volume of information run interference? Or simply, could there be times when a high volume of incorrect information 'wins' over a low volume of 'more correct' information? If so, how should this be handled? In our broad project work, we are attempting to optimise interfaces to ensure that trustworthiness and the risk of trusting certain information is properly communicated [20]. Therefore, we will need to combat volume/quantity interference across all operational contexts.

ACKNOWLEDGMENT

This work was conducted as a part of the TEASE project, a collaboration between the University of Oxford, 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] B. McGuinness and A. Leggatt, "Information trust and distrust in a sensemaking task," in *Command and Control Research and Technology Symposium*, 2006.

[2] J. R. C. Nurse, S. S. Rahman, S. Creese, M. Goldsmith, and K. Lamberts, "Information quality and trustworthiness: A topical state-of-the-art review," in *International Conference on Computer Applications and Network Security*. IEEE, 2011, pp. 492–500.

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

[4] E. Sillence, P. Briggs, P. Harris, and L. Fishwick, "A framework for understanding trust factors in web-based health advice," *International Journal of Human-Computer Studies*, vol. 64, no. 8, pp. 697–713, 2006.

[5] J. Edelenbos and E. Klijn, "Trust in complex decision-making networks," *Administration & Society*, vol. 39, no. 1, pp. 25–50, 2007.

[6] A. Leggatt and B. McGuinness, "Factors influencing information trust and distrust in a sensemaking task," in *11th International Command and Control Research and Technology Symposium*, 2006.

[7] A. Pickard, P. Gannon-Leary, and L. Coventry, "Trust in 'E': Users' trust in information resources in the web environment," *Enterprise Information Systems*, pp. 305–314, 2010.

[8] Y. Gil and D. Artz, "Towards content trust of web resources," *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, no. 4, pp. 227–239, 2007.

[9] A. Patrick, S. Marsh, and P. Briggs, "Designing systems that people will trust," National Research Council Canada, Tech. Rep., 2005.

[10] E. Costante, J. den Hartog, and M. Petkovic, "On-line trust perception: What really matters," in *1st Workshop on Socio-Technical Aspects in Security and Trust (STAST) at 5th International Conference on Network and System Security (NSS)*. IEEE, 2011, pp. 52–59.

[11] S. Moturu and H. Liu, "Quantifying the trustworthiness of social media content," *Distributed and Parallel Databases*, vol. 29, no. 3, pp. 239–260, 2011.

[12] N. Idris, M. Jackson, and R. Abrahart, "Colour coded traffic light labeling: A visual quality indicator to communicate credibility in map mash-up applications," in *International Conference on Humanities, Social Sciences, Science & Technology*, 2011.

[13] T. Lucassen and J. Schraagen, "Evaluating wikitrust: A trust support tool for Wikipedia," *First Monday*, vol. 16, no. 5, 2011.

[14] Wikipedia, "Wikipedia:article feedback tool," 2012. [Online]. Available: http://en.wikipedia.org/wiki/Wikipedia:Article_Feedback_Tool

[15] A. Pang, C. Wittenbrink, and S. Lodha, "Approaches to uncertainty visualization," *The Visual Computer*, vol. 13, no. 8, pp. 370–390, 1997.

[16] A. Bisantz, R. Stone, J. Pfautz, A. Fouse, M. Farry, E. Roth, A. Nagy, and G. Thomas, "Visual representations of meta-information," *Cognitive Engineering & Decision Making Journal*, vol. 3, no. 1, pp. 67–91, 2009.

[17] A. Bisantz, D. Cao, M. Jenkins, P. Pennathur, M. Farry, E. Roth, S. Potter, and J. Pfautz, "Comparing uncertainty visualizations for a dynamic decision-making task," *Cognitive Engineering & Decision Making Journal*, vol. 5, no. 3, pp. 277–293, 2011.

[18] D. Battre, K. Djemame, O. Kao, and K. Voss, "Gaining users' trust by publishing failure probabilities," in *3rd Conference on Security and Privacy in Communications Networks*. IEEE, 2007, pp. 193–198.

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

[20] J. R. C. Nurse, S. Creese, M. Goldsmith, and K. Lamberts, "Trustworthy and effective communication of cybersecurity risks: A review," in *1st Workshop on Socio-Technical Aspects in Security and Trust (STAST) at 5th International Conference on Network and System Security (NSS)*. IEEE, 2011, pp. 60–68.