Trust and Reliance Based on System Accuracy

Kun Yu*
Shlomo Berkovsky
Dan Conway
Ronnie Taib
Jianlong Zhou
Fang Chen
Data61, CSIRO
13 Garden Street, Eveleigh, NSW 2015, Australia
{firstname.lastname}@data61.csiro.au

ABSTRACT
Trust plays an important role in various user-facing systems and applications. It is particularly important in the context of decision support systems, where the system's output serves as one of the inputs for the users' decision making processes. In this work, we study the dynamics of explicit and implicit user trust in a simulated automated quality monitoring system, as a function of the system accuracy. We establish that users correctly perceive the accuracy of the system and adjust their trust accordingly.

Keywords
User-system trust; system accuracy; trust formation; reliance

1. INTRODUCTION
User-system trust is an important construct in human-computer interaction as well as in many practical user-facing systems. It is particularly important for systems where users are required to make decisions based, at least partially, on machine recommendations. For instance, consider a medical decision support system or an e-commerce recommender system. In both, a user decides on the course of actions – be it medical treatment for a patient or product to purchase – in uncertain conditions and based (in part) on the system's suggestions. Due to the possible negative implications of incorrect decisions, the lack of user trust may deter the user from following these suggestions and be detrimental to the acceptance of system recommendations.

Hence, trust in automation, and in particular decision support information technologies, has been the focus of many studies over the last decades [1], [2]. It has mainly been studied in the context of task automation and industrial machinery. In one of the seminal works in this field, Muir et al [3] found a positive correlation between the level of user trust and the degree to which the user delegated control to the system. Furthermore, McGuirl and Sarter [4] found similar responses specifically within an automated decision support system. Note that both works measured the impact of establishing and maintaining trust based on user reliance on system suggestions, indirectly deriving the uptake of the system.

Although much work has been devoted to the impact of system performance [5] and transparency [6] on user trust, less attention has been paid to the temporal variations of trust. In this work, we set out to investigate the fine-grained dynamics of trust in an experiment that simulates an Automated Quality Monitoring (AQM) system that alerts users to of the existence of faulty items, in a fictional factory production line scenario. In the experiment, 22 participants interacted with four AQM systems, each exhibiting a different level of accuracy. After each trial (30 per AQM system), the users reported their perceived level of trust in the system, which we refer to as *explicit trust*. In addition, we also measured *implicit trust* through reliance, quantified by the portion of times the user has followed the AQM's suggestions so far.

Two hypotheses guided our work:
- H1: Trust would stabilise over time to a level correlated with the system's accuracy;
- H2: Users would exhibit thresholds of acceptable accuracy for the system, under which reliance would drop.

This work experimentally validates these hypotheses and draws practical conclusions that can help system designers maintain user trust in systems. In the following sections, we first present related work on user-system trust, followed by a detailed description of the experimental protocol. We then present and discuss the results, and finally conclude with a discussion on practical steps that might be taken to sustain user trust.

2. RELATED WORK
Human-machine trust has generated an extensive body of literature since it was originally investigated within the context of industrial automation systems in the nineties. Although multiple definitions, frameworks and decompositions of trust exist, there is convergent evidence about its central characteristics. We adopt the definition proposed by Lee and See [7] where '*trust can be defined as the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability.*' This succinctly encapsulates the primary sources of variance (the user, the system, the context) and identifies a key aspect of this relationship – that of vulnerability. Similar definitions exist by Rousseau et al. [8], Mayer et al. [9] and Hoff and Bashit [2]. Trust is a hypothesised variable that has been shown to be a key mitigating factor in system use/disuse (i.e., reliance) [1]. It can be inferred from both self-reported and behavioural measures [10], and importantly, is dynamic, with acquisition and extinction curves, subject to the user's perception of system performance.

Trust has also been proposed to be a multi-dimensional construct with a number of models existing in the current literature, each with slightly different proposed component subscales. We have adopted the model of [2], which was based on an empirical research overview of existing literature. This model proposes that three conceptual types of factors influence user-system trust. *Dispositional* trust reflects the user's natural tendency to trust machines and encompasses cultural, demographic, and personality factors. *Situational* trust refers to more specific factors, such as the task to be performed, the complexity and type of system, user's workload, perceived risks and benefits, and even mood. Lastly, *learned* trust encapsulates the experiential aspects of the construct, which are directly related to the system itself. This

UMAP '16, July 13-17, 2016, Halifax, NS, Canada
© 2016 ACM. ISBN 978-1-4503-4370-1/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2930238.2930290

variable is further decomposed into two components. One is *initial learned* trust, which consists of any knowledge of the system acquired before interaction, such as reputation or brand awareness. This initial state of learnt trust is then also affected by *dynamic learned trust*, which develops as the user interacts with the system and begins to develop experiential knowledge of its performance characteristics such as reliability, predictability, and usefulness. The relationships and interplay between these factors influencing trust are complicated and subject to much discussion. In our work we focussed on how trust changes through human-machine interaction and, therefore, seek to manipulate experimental variables thought to influence dynamic learned trust whilst keeping situational and dispositional variables static.

3. METHODOLOGY

3.1 Context

We operationalised a binary decision making task in our experiment for two reasons. Firstly, any complex decision process can be arguably decomposed into a series of binary decisions. The decision-trust relationship, thus, can be easily generalised into complicated decision-making problems. Secondly, the simplified decision making protocol we implemented, similar in effect to the 'micro-worlds' discussed by Lee and See [7], makes it convenient to map trust to decisions without other parameters' interference.

The scenario of the experiment was a typical industrial quality control task. This simulated task consisted of checking the quality of drinking glasses on a production line, with the assistance of a decision support system called an *Automatic Quality Monitor* (AQM). However, the AQM was not always correct, i.e., it would occasionally exhibit false positives (suggesting failing a good glass) and misses (suggesting passing a faulty glass).

3.2 Trials

Each trial required the participant to make a decision about whether to pass or fail a glass, with no other information about the glass other than the AQM's suggestion. Trials were presented sequentially, providing a time-based history of interaction with a given AQM. In each trial, the participant could trust the AQM or override it and make his/her own decision.

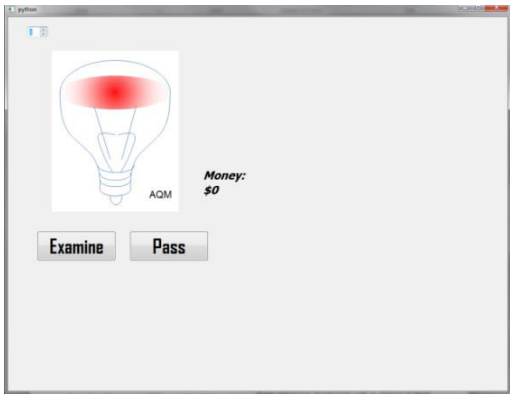

Figure 1: The trial starts with an AQM suggestion.

Each trial starts with the AQM providing a suggestion for a new glass as shown in Figure 1, by illuminating a red warning light-bulb if it predicts the glass to be faulty. Otherwise the warning light remains off. It should be noted that the status of the AQM light and the possible quality of the glass are both binary features to help generalise results, as mentioned above. The participant must then decide whether to pass the glass by clicking the *Pass* button, or conversely to fail the glass by clicking the *Examine* button. The actual glass is then displayed, so the participant receives direct feedback on their decision, as shown in Figure 2.

Furthermore, we gamified the experiment in an attempt to increase motivation and attention: each time the participant made a correct decision, i.e., examined a faulty glass or passed a good glass, they earned a fictional $100 reward. However, each incorrect decision cost them a fictional $100 fine. The total earnings were updated and displayed after each decision. Note that the rewards and the fines were used for gamification purposes only, and no actual remuneration was offered to the participants.

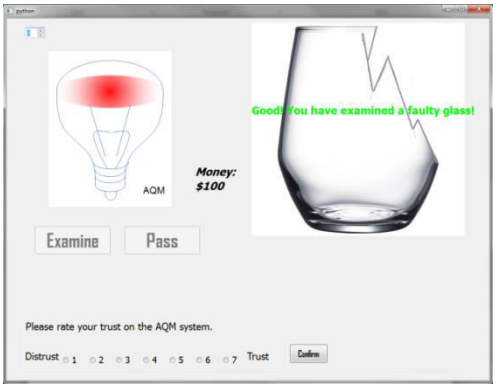

Figure 2: Upon the participant's decision, the actual glass condition is shown and score is updated.

3.3 AQM Accuracy and Blocks

The experiment session was separated into four blocks, and participants were instructed that a different AQM was used for each block. The accuracy of each of the four AQMs presented was manipulated by varying the average rate of false positives and false negatives exhibited by each system. These errors were presented in a randomised order within the 30 trials presented for each participants and each AQM.

We used four different AQM accuracies, i.e. 100%, 90%, 80% and 70% respectively. In order to capture a trust baseline for each participant, each experiment session always started with the 100% accuracy AQM, followed by the other three AQMs used in a random order. Each AQM was used for 30 task trials. The AQM made errors randomly over the trials, but in a way that the mean accuracy for respective AQM was as defined. For instance, the 80% AQM would make, on average, 6 errors over the 30 trials (on average, 3 false positives and 3 false negatives).

3.4 Participants and Data Collection

Twenty-two participants took part in the 45 minute experiment. The participants were university students and IT professionals. No specific background or requirements were required to complete the task. Recruitment and participation were conducted in accordance to an approved ethics plan for this study. No reward or compensation was offered for taking part in the experiment.

For each trial, we collected:

- The participant's binary decision (pass or examine);
- The AQM suggestion (light on or light off);
- The actual glass condition (good or faulty);
- The subjective trust rating, collected after the actual state of the glass had been revealed. This rating was collected using a 7-point Likert scale ranging from 1=distrust to 7=trust. In the instructions issued at the outset of the experiment we explained that a rating of 4 meant neutral, or no disposition in either direction.

One of the participants had consistently rated the trust at extreme levels (either 1 or 7) of the 7-point scale across the four AQMs, hence, their data was excluded from the analysis. Considering individual differences, the trust data was normalised to the [0,1] range on individual basis, for all the trials conducted on the four AQMs. The binary decisions of the participants were further quantified in terms of a reliance score, i.e., ratio between the number of decisions consistent with the AQM suggestions and the total decisions for a set number of consecutive trials.

4. RESULTS

In this section we present and discuss the results of our examinations on trust in the light of our hypotheses.

4.1 Trust Correlation to System Accuracy

We start with the investigation of acquisition and extinction of trust, as observed over the course of user interactions with the AQMs. The level of trust is measured subjectively after each trial, as described earlier. Since the AQM errors were randomised over the 30 trials for each AQM, and given the small number of participants, trust variations for each trial exhibit a number of local fluctuations. We address this by applying a simple low-pass filter; specifically, a 5-trial sliding window, reducing our data to 25 points per AQM. That is, the level of trust after trial *N* was computed as the average trust across the last 5 trials. Figure 3 shows the aggregated normalised trust scores for all 21 participants.

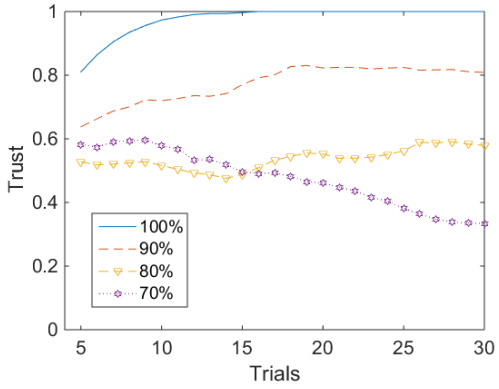

Figure 3: Mean trust for all participants, all AQMs.

Initially, trust in the 90%, 80%, and 70% AQMs seems comparable, as would be expected, since participants know that each new AQM is different from the others they may have encountered, and the order is randomised. However, the initial trust in the 100% AQM appears to be above the other AQMs, as it was the first AQM that users interacted with.

An analysis of variance showed that the effect of the AQM accuracy on the first trust point (trust mean over the first five trials) was significant for all participants, $F(3, 80)=6.463$, $p<0.001$. Post hoc Tuckey tests show there is a significant difference between the 100% AQM and both the 80% and 70% AQMs. We hypothesise that this may be linked to the sliding window we used to capture trust, as the participants started to form a preliminary trust judgment of each AQM by the time of the first trust point (recall that the first point is actually after 5 trials).

As a side note, the test of homogeneity (Levene's) for the first reliance point was significant, hence violating ANOVA's assumption of equal variances. However, the sample sizes being equal, this statistic should be robust. Hence, we accept the results.

Looking at the temporal fluctuations of the trust values, we observe that these stabilise with important differences observed between the AQMs. As expected, trust in the 100% AQM stabilises at 1 after 13 trials only. Also the 90% AQM converges to reasonably high levels of trust from trial 19. The 80% AQM is initially stable but exhibits a slight increase in trust starting from trial 15, while the trust in the 70% AQM steadily declines after less than 10 trials and eventually drops as low as 0.33.

An analysis of variance showed that the effect of the AQM accuracy on the last trust point (trust mean over the last five trials) was significant for all the participants, $F(3, 80)=27.03$, $p<0.001$. Post hoc Tuckey tests show there is a significant difference between the 100% AQM and both the 80% and 70% AQMs, as well as between the 90% AQM and the 70% AQM, and, again, between the 80% AQM and the 70% AQM.

It should be noted that the final order of the trust ratings corresponds to that of the AQM accuracies. That is, the 100% AQM stabilises at the highest trust level, followed by the 90% AQM, 80% AQM, and 70% AQM, in this order. This finding supports our hypothesis H1 that *trust would stabilise over time to a level correlated with the systems' accuracy*.

In addition, since only a small set of discrete accuracies were examined, it can be interesting to analyse our results from a rank-ordering problem perspective. Indeed, this would provide an indication of whether the reported trust ranking aligns to the discrete accuracy levels of the AQMs. A Friedman's test shows significant differences between the trust levels (Friedman's $\chi 2$ (20, 3) = 45.31, $p < 0.001$), with mean ranks of 3.8, 2.9, 2.0 and 1.3 for the 100%, 90%, 80% and 70% AQMs, respectively. These statistics suggest that trust ratings correlate with increased levels of AQM accuracy, when considered as discrete values (here, 10% increments), again, supporting our hypothesis H1.

4.2 Acceptable Accuracy and Reliance

The dynamics of reliance are regarded as an objective measure of trust. Recall that reliance was measured implicitly during each trial as described earlier. Again, we applied in this case a simple low-pass filter, but this time we used a 10-trial sliding window, reducing our data to 20 points per AQM. The reason for this larger window is mainly because reliance is a binary feature (at every trial the participant either did or did not follow the AQM's suggestion). Hence, local variations tend to add weight to the reading for a small window size. Figure 4 shows the aggregated reliance for all the 21 participants and all four AQMs.

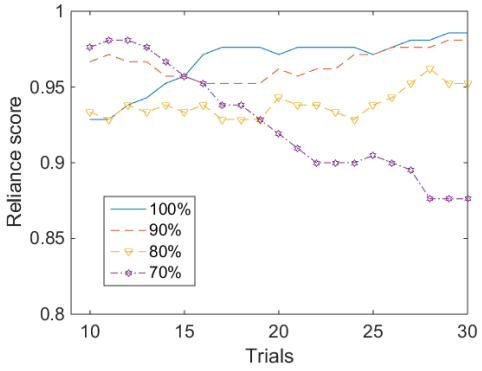

Figure 4: Mean reliance for all participants, all AQMs.

We observe that despite the twice larger sliding window, the reliance curves are less stable than the trust curves. We believe that the reason for this observation is two-fold. Firstly, the effect of a binary feature on smoothing is strong and could require a

wider sliding window, but this would mean losing temporal accuracy in our analysis of reliance dynamics. Secondly, we posit that while participants exhibit relatively uniform trust trends, they have different strategies to deal with it and this comes through the objectively measured reliance values.

All curves, except for the 70% AQM, demonstrate slight (and often unstable) increases and their final levels are in the range of 0.95-0.98. The 100% and 90% AQMs seem to converge strongly, while the 70% AQM exhibits a steady decline in reliance. The 80% AQM seems close to the 100% AQM baseline. This could indicate that the acceptable level of accuracy for a system is around the 80% mark, since the reliance of the 80% AQM is slightly lower.

An analysis of variance showed that the effect of the AQM accuracy on the first reliance point was not significant for all participants, *F(3, 80)=01.597, p=0.197 n.s*. That is, the apparent reliance pairs observed are not significant in view of the variance, further demonstrated by Figure 5. This means that the participants interacted with all four AQMs with a comparable level of dispositional trust, as indicated by the implicit reliance measure.

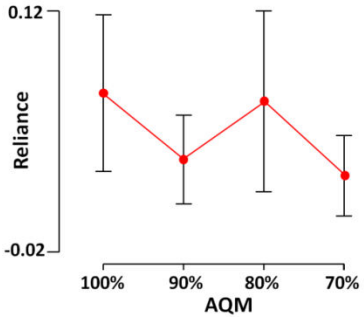

Figure 5: First reliance point variance for all participants.

Focusing on the last reliance observed after 30 trials, an analysis of variance showed that the effect of the AQM accuracy on the last reliance point was significant for all participants, *F(3, 80)=4.182, p=0.008*. The test of homogeneity (Levene's) was significant, but, again, the sample sizes are equal.

Due to the binary notion of reliance, we can test our hypothesis of acceptable level of accuracy by comparing all the AQMs to the 100% AQM baseline, in order to determine where the threshold for accuracy may lay. To do so, we applied a simple contrast in the ANOVA for the last reliance point, and obtained significance only for the pair 100% AQM versus 70% AQM. This means that the 80% AQM, while being visually apart from the 100% and 90% AQMs, is actually not significantly different. However, the 70% AQM is significantly different from the other three AQMs.

These results support our hypothesis H2 that *users have thresholds of acceptable accuracy for a system, under which the reliance drops*. Since there is no significant difference between the AQMs in terms of the initial reliance levels, participants start interacting with the AQMs free of pre-disposition. But later on we observed a different behaviour only for the 70% AQM, whereby the reliance of the participants on that AQM declined significantly compared to other AQMs. This indicates that a threshold of acceptable accuracy in the AQMs, as observed from our participants, lies somewhere between the 70% and 80% levels.

Having said that, the high values and narrow range of reliance values should be highlighted. Over the course of the whole experiment, reliance curves of all the four AQMs remain fairly compact and above the 0.9 mark. This behaviour is not surprising, however, and can be explained by the relatively high accuracies exhibited by all the AQMs. Even the poorest AQM operating at the 70% accuracy can correctly monitor the quality of a glass in 7 cases out of 10, which is well above chance. We hypothesise that the participants rightfully perceived this benefit of the AQM over pure random choice. Hence, they followed the AQM's suggestions, leading to very high levels of reliance.

5. DISCUSSION AND CONCLUSION

In this work, we investigated the fine-gained dynamics of user-system trust, an important construct of human interaction with a decision support system. We specifically focused on an automated quality monitor (AQM) simulation, which provided indication of faulty glasses being produced. In our study, each user interacted with four AQMs and out of these interactions we populated the explicit trust and implicit reliance scores.

We analysed the temporal dynamics of both trust and reliance, as well as their dependence on the accuracy exhibited by the AQM. It was found that the reported trust levels aggregated across the entire cohort of users, stabilised over time and, in general, corresponded to the accuracy of the AQMs. Somewhat surprisingly, we discovered that the implicit reliance levels were very high and comparable across the four AQMs. We attribute this finding to the relatively high accuracy levels of the AQMs in our experiment.

Hence, the obtained experimental results support the hypotheses raised at the beginning of the paper. Firstly, we observe that the learned user-system trust stabilised over time and generally correlated with the level of accuracy exhibited by the system. Secondly, our findings indicate that at reasonably high levels of system accuracy, user reliance is high, whereas once the system accuracy falls below an acceptance threshold, the reliance is likely to deteriorate as well.

It should be noted that our findings are based on a reasonably limited cohort of participants, all having reasonably short interactions with the system. In the future, we would like to increase the number of interactions so that we may reduce the frequency of users reporting their explicit trust. For example, we could collect the explicit trust level every second interaction, allowing us to double the length of interactions without over-burdening the users. This would allow us to collect more solid empirical evidence and better support our hypotheses.

Finally, more work is needed to address the fine-grained dynamics of trust acquisition and extinction. In our work, we assumed a stable level of accuracy of every system. This, however, may vary over the course of user interaction. Hence, it is important to validate the evolution of user trust as a function of the user's initial trust disposition, observed system performance, and temporal aspects of this performance (e.g., initial failures vs. failures when the trust was already formed). Furthermore, it is possible to depict different user's trust evolving path according to their decision making patterns, and hence users can be categorized in the light of their respective trust profiles. We highlight the importance of these research questions, but leave this work for the future with expanded collection of user trust and interaction data.

ACKNOWLEDGMENTS

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. This work was also partially supported by the Asian Office of Aerospace Research & Development (AOARD) under grant No. FA2386-14-1-0022 AOARD 134131.

REFERENCES

[1] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *Int. J. Hum.-Comput. Stud.*, vol. 40, no. 1, pp. 153–184, Jan. 1994.

[2] K. A. Hoff and M. Bashir, "Trust in Automation Integrating Empirical Evidence on Factors That Influence Trust," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 57, no. 3, pp. 407–434, May 2015.

[3] B. M. Muir, "Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, Nov. 1994.

[4] J. M. McGuirl and N. B. Sarter, "Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 48, no. 4, pp. 656–665, Dec. 2006.

[5] W. Wang and I. Benbasat, "Attributions of Trust in Decision Support Technologies: A Study of Recommendation Agents for E-Commerce," *J. Manag. Inf. Syst.*, vol. 24, no. 4, pp. 249–273, Apr. 2008.

[6] J. Zhou, "Transparent Machine Learning—Revealing Internal States of Machine Learning," in *Proceedings of IUI2013 Workshop on Interactive Machine Learning*, Santa Monica, CA, 2013.

[7] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 46, no. 1, pp. 50–80, Mar. 2004.

[8] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not So Different After All: A Cross-Discipline View Of Trust," *Acad. Manage. Rev.*, vol. 23, no. 3, pp. 393–404, Jul. 1998.

[9] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model Of Organizational Trust," *Acad. Manage. Rev.*, vol. 20, no. 3, pp. 709–734, Jul. 1995.

[10] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an Empirically Determined Scale of Trust in Automated Systems," *Int. J. Cogn. Ergon.*, vol. 4, no. 1, pp. 53–71, Mar. 2000.

[11] J. Rotter, "An new scale for the measurement of interpersonal trust," *J. Pers.*, vol. 35, no. 4, pp. 651–665, 1967.

[12] C. L. Scott, "Interpersonal Trust: A Comparison of Attitudinal and Situational Factors," *Hum. Relat.*, vol. 33, no. 11, pp. 805–812, Nov. 1980.

[13] I. L. Singh, R. Molloy, and R. Parasuraman, "Automation-Induced 'Complacency': Development of the Complacency-Potential Rating Scale," *Int. J. Aviat. Psychol.*, vol. 3, no. 2, pp. 111–122, Apr. 1993.

[14] P. Madhavan and D. A. Wiegmann, "Similarities and differences between human–human and human–automation trust: an integrative review," *Theor. Issues Ergon. Sci.*, vol. 8, no. 4, pp. 277–301, Jul. 2007.

[15] P. C. Earley, "Computer-generated performance feedback in the magazine-subscription industry," *Organ. Behav. Hum. Decis. Process.*, vol. 41, no. 1, pp. 50–64, Feb. 1988.

[16] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *J. Exp. Psychol. Gen.*, vol. 144, no. 1, pp. 114–126, 2015.

[17] M. T. Dzindolet, L. G. Pierce, H. P. Beck, and L. A. Dawe, "The Perceived Utility of Human and Automated Aids in a Visual Detection Task," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 44, no. 1, pp. 79–94, Mar. 2002.

[18] S. Berkovsky, J. Freyne, and H. Oinas-Kukkonen, Eds., "Influencing Individually: Fusing Personalization and Persuasion," *ACM Trans Interact Intell Syst*, vol. 2, no. 2, pp. 9:1–9:8, Jun. 2012.

[19] S. Merritt and D. Ilgen, "Not all trust is created equal: Dispositional and history-based trust in human-automation interactions," *Hum. Factors*, vol. 50, no. 2, pp. 194–210, 2008.