How the Accuracy of Interactive Voice Assistants Affect Perceived Trust

Wenhu Zhang and Chris Baber

School of Computer Science, University of Birmingham

ABSTRACT

We asked 30 participants to ask questions of an Interactive Voice Assistant (IVA) which we had modified to provide different levels of accuracy in its answers. The levels of accuracy were low (55%) or high (80%). We also told users what level of accuracy to expect (60% or 100%). This produced a set of 6 combinations of actual accuracy with expected accuracy (including the condition when we did not tell the users which level of accuracy to expect). As expected, when users experience a more reliable IVA (i.e., 80% vs. 55%) their rating of trust is higher, and when actual an IVA with high accuracy and they are expecting accuracy to be high, then their trust rating is higher still. However, *expected* accuracy seems to outweigh actual accuracy, particularly when the actual performance is less than expected. Counter intuitively, this suggests that participants were not able to judge the actual accuracy of the IVA but relied on the *expected* accuracy.

KEYWORDS

Trust, Interactive Voice Assistants

Introduction

The aim of this study was to explore how Trust if affected by accuracy when people speak to an Interactive Voice Assistant. Specifically, we were interested in how the level of accuracy that users expect from the device compares with the level of accuracy that they experience when using the device. The point of manipulating expected accuracy (in addition to the actual accuracy) was to explore interaction effects on trust arising from expected and actual accuracy. For example, if users expect the device to have low accuracy and it performs very well, does this have a positive impact on their trust in the device.

Research into human trust in automation (TiA)began in earnest with pioneering work of Muir (1994) and Lee and Moray (1992). When it comes to perceiving the level of trust in an automation, users have few criteria to judge it other than its stated performance, their observations during actual use and their own knowledge domain. This raises the question of whether users can determine the reliability or the accuracy of the automation with which they are interacting, or whether they rely on prior experience of their interactions, or simply revert to trusting performance claims provided to them. For example, would people trust an automation more if they were told that it was 90% accurate on retained data rather than 60%? If more trusting, does its stated accuracy when they actually use the automation still affect that trust?

There remains a lack of a universally accepted model of TiA, perhaps because trust varies with context and type of automation. To some extent, this assumes that 'trust' is dispositional, i.e., trust is a subjective response to the performance of automation. In this paper, we adopt Mayer and Davis's (1995) model of trust (figure 1). In this model, trust arises from a combination of the user's propensity to trust and factors which affect perceived trustworthiness of the agent (human or

automation) they interact with. The experience of using automation could lead to changes in user response (either physiologically or behaviourally), particularly when the agent behaves in unexpected ways. This implies that, while the model appears to be dispositional, it has a strong activity component.

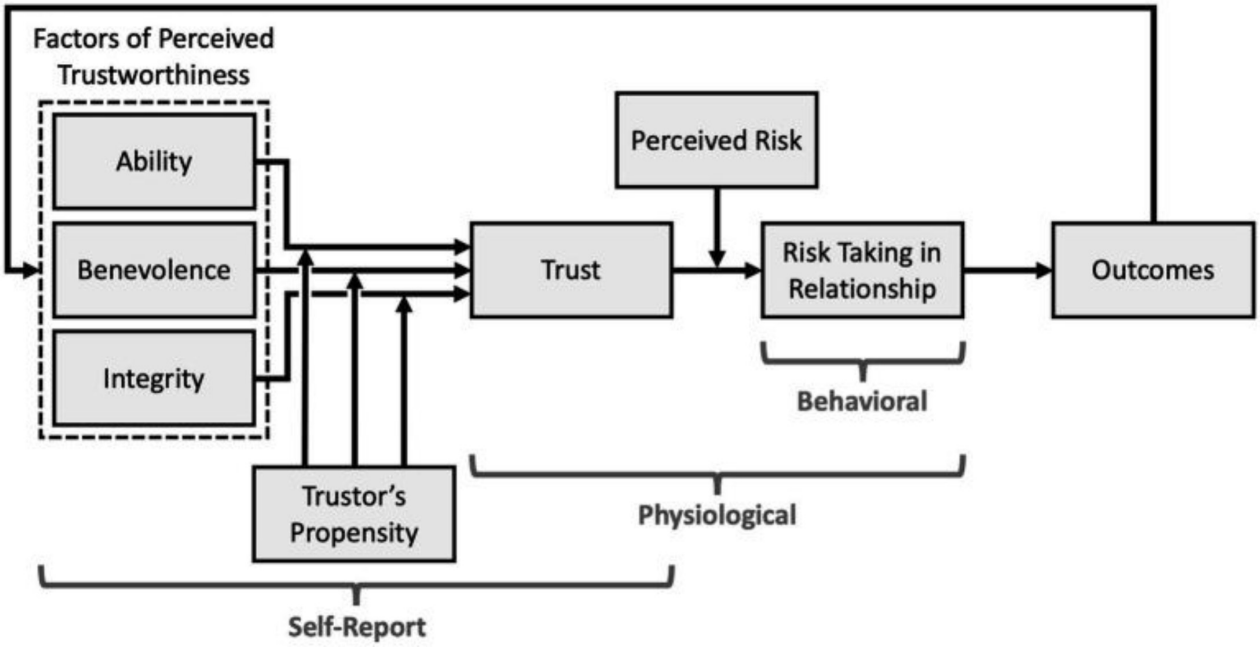

Figure 1: Mayer and Davis' (1995) model of trust

Kohn et al. (2021) consider Meyer and Davis' (1995) model in terms of the ways in which trust could be measured (table 1).

Table 1: Relating types of trust to measures

Trust Type	Trust Process Step	Measure	Experiment step
Factors of perceived trustworthiness	Perception of the system's trustworthiness-related characteristics	Self-report from user	Before / during interaction
Trustor's propensity	Effects of individual's traits	Self-report from user	Before interaction
Trust	Trust stance or attitude that exists during interactions and influenced by feedback	Self-report; Physiological measures	During / after interaction
Perceived risk	Effects of individual's understanding of situation	N/A	Pre-interaction in environmental situation
Risk-taking	Trust behaviour expressed during interactions	Behaviour	During / after interaction
Outcomes	System accuracy and user trust	N/A	After interaction

Method

Participants

We recruited 30 participants for this study (mean age 26 (± 15); 9 female). Participants had no previous experience of using Interactive Voice Assistants.

Equipment

We used Google's Dialogflow to create a bespoke Interactive Voice Assistant (IVA) and participants interacted with a Bose Soundlink Revolve II (figure 2) to present spoken response to questions.

Figure 2: Bose Soundlink revolve II

We modified the performance of the IVA so that it provided correct answers to either 55% or 80% of the questions. The correct was specified in our question set; an incorrect answer was a random choice of answer from the question set.

Measures

We used Jian et al's. (2000) Human-Automation Trust Checklist to gather subjective responses from participants. This is a scale that has been widely used in the literature. This checklist includes 12 questions which reflect a range of attitudes which can affect the perception of trust in automation. The checklist has separate sets of questions for Distrust and Trust (5 for distrust and 7 for trust). While it is popular means of evaluating trust, Gutzwiller et al. (2019) urge caution in its use because it could be skewed towards positive ratings. When the survey is completed in the original order of statements and with the original rating scales used, participants tended to produce higher ratings of trust than when the survey was presented in other configurations (albeit the effect was quite small). As a simple expedient to minimise the potential bias, we subtracted the median ratings for the Trust questions from those of the Distrust questions (on the assumption that this was indicate propensity to trust the IVA). However, this means that the application of the Human-Automation Trust Checklist is different from the original intentions of its developers.

Procedure

We assumed that accuracy would primarily be affected by the IVA misrecognising a complete sentence, rather than individual words. In this manner, we could compare the impact of medium and high accuracy of user trust. Additionally, we told participants that the IVA had an accuracy of either 60% or 100%.

The combination of experienced and expected accuracy produced six experimental conditions, as shown in table 1. The order in which participants experienced the conditions was counter-balanced using Latin squares. We repeated the conditions for each question set. Examples of questions are shown in table 1.

Table 1: Examples of questions used for the experiment

Question Set1	**Answer**
What is 20% of 80?	Twenty percent of eighty is sixteen
What does "Sociable"mean?	Sociable means to have a harmonious relationship with everyone and get along well
How many days are there in a year?	There are three hundred and sixty-five days in a year
How many kilometres are in a mile?	A mile is about 1.6 kilometres
How is the road to work?	Good road to work
How much is five plus seven?	Five plus seven equals twelve
What day is it today?	Today Thursday
How is the traffic situation?	Good traffic conditions nearby
How's the weather today?	It's sunny now, the temperature is 30 degrees
Question Set2	**Answer**
Add meeting to my calendar	All meetings added to calendar
Adjust the temperature of the bedroom air conditioner to 26 degrees	The bedroom air conditioner has been adjusted to 26 degrees
Add milk to my shopping list	Added milk to shopping list
What is the current volume?	The current volume is 50%
When is my first meeting today?	First meeting today at 3pm
What questions can I ask you?	You can ask me for help with information and daily tasks.
What is the battery level of my speaker?	Battery is 66%
Remind me to call mom every Sunday	OK, the reminder is set to start this Sunday
When will sunrise tomorrow?	Sunrise tomorrow is 5:30
How is the S&P 500 performing?	S&P 500 shares rose to 3998.95 today, up 0.99%

Results

Initial analysis of the responses to the checklist was performed, for each question set and across each experimental condition, using Cronbach's alpha and Kaiser-Meyer-Olkin sampling adequacy. The results, shown in table 2, indicate high levels of agreement within the checklists across all conditions. From this, we assumed that it would be appropriate to merge responses to question sets 1 and 2 for subsequent analysis.

Table 2: Agreement of participants in their trust ratings

Accuracy (expected)	60%	100%	60%	100%	Not told	Not told
Accuracy(actual)	55%	55%	80%	80%	55%	80%
Condition	A	B	C	D	E	F
Cronbach's alpha	0.847	0.8295	0.873	0.8325	0.855	0.847
KMO	0.7675	0.7965	0.8305	0.702	0.7705	0.7675

We analysed the median rating of the seven 'Trust' questions (figure 1). A Friedman Analysis of Variance, calculated using R, showed a significant main effect of Condition [x2 (5) = 109.5, p<0.0001]. With the exception of C x F, all post-hoc comparisons (using Wilcoxon Signed Ranks test) were significant at the 5% level. The highest rating of trust was for condition D (in which expected and actual accuracy were high).

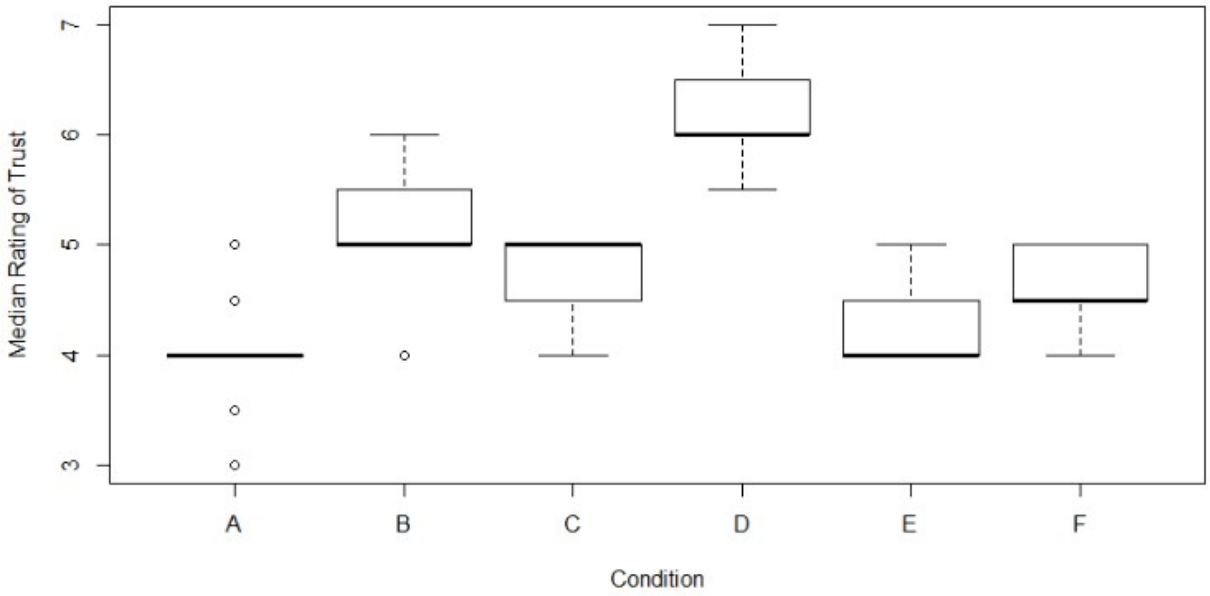

Figure 1: Rating of 'Trust' between the different conditions

Conclusions

We compared trust ratings when using an IVA under different manipulations. We were interested in how expected accuracy or actual accuracy affected these ratings. The results in figure 1 can be grouped into four observations.

First, conditions A (low expected + low actual accuracy) and E (no expected + low actual accuracy) are similar. This suggests that participant could detect when the actual accuracy of the IVA was low.

Second, condition D (high expected + high actual accuracy) is significantly different to the other conditions (at p<0.05 using Wilcoxon pairwise, post-hoc tests). This suggests that participants were positively influenced by high expected and high actual accuracy. While this is to be expected, it suggests that (coupled with observation 1) that participants were moderating their trust ratings in a predictable manner.

Third, when participants have not been told the accuracy of the IVA (conditions E and F), there is no significant difference in trust in terms of actual accuracy. This was surprising, given observation 1, because it suggests that detection of actual accuracy is not as simple as we might assume. It might be that we had too few questions in our question sets so that participants did not have long

enough exposure to the IVA to form an opinion of its accuracy. An alternative explanation that, without being given expected accuracy, participants begin with a low expected accuracy (possibly lower than the one we provided) because they are not sure whether the IVA would recognise their speech. This was mentioned by a few of the participants. In this case, rather than the 'trust' being in the IVA it would be based on whether the IVA would respond to the participants (and, the implication here is that the participants might place the locus of any performance failures on themselves and their inability to get the IVA to work as much as on the failure of the IVA to respond to them).

Fourth, the trust rating conditions B (high expected + low actual accuracy), C (low expected + low actual accuracy), and F (no expected + high actual accuracy) show no difference. This also suggests that rating of trust is moderated by actual accuracy (B and C) but that there is also an a priori assumption that the IVA will have low performance.

Our findings bring a number of practical implications for research on Human-Computer interaction and trust in automation, thereby increasing trust in automation. Firstly, our findings show that designers of automation must express their expectations of automation accurately and responsibly, because only in this way can users determine the extent to which they can trust automation before and after interaction. Furthermore, we have found that users focus on the process of interacting with automation even when their interaction with it is limited to a few tasks. For example, if the actual use is very different from the automation training, i.e., if the stated accuracy does not reflect well the accuracy observed in its actual use. However, this small amount of task feedback is not indicative of the average performance of the automation in actual use. Therefore, it is important for automation designers to communicate well the uncertainty that the automation will complete correctly based on a small number of tasks. In this way, even if the user observes low performance on the successful completion of the first few tasks when using the automation, this will not lead to a false distrust of an automation. All of the above are implications of the project's findings for automated trust research.

Finally, our work highlights that in order to understand the interactions between humans and the different components of automated work, more experimentation is needed to enable work on the interpretable aspects of machine learning within automation to go beyond the current focus on its models themselves.

References

Gutzwiller, R.S., Chiou, E.K., Craig, S.D., Lewis, C.M., Lematta, G.J. and Hsiung, C.P., 2019, November. Positive bias in the 'Trust in Automated Systems Survey'? An examination of the Jian et al.(2000) scale. In *Proceedings of the Human Factors and Ergonomics Society annual meeting*(Vol. 63, No. 1, pp. 217-221). Sage CA: Los Angeles, CA: SAGE Publications.

Jian, J., Bisantz, A., & Drury, C. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

Kohn, S.C., de Visser, E.J., Wiese, E., Lee, Y.C. and Shaw, T.H., 2021. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in psychology*, *12*, p.604977.

Lee, J. and Moray, N., 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), pp.1243-1270.

Mayer, R.C., Davis, J.H. and Schoorman, F.D., 1995. An integrative model of organizational trust. *Academy of management review*, *20*(3), pp.709-734.

Muir, B.M., 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), pp.1905-1922.

Appendices A Jiang et al. Human-Automation Trust Checklist

Checklist for Trust between People and Automation

Below is a list of statement for evaluating trust between people and automation. There are several scales for you to rate intensity of your feeling of trust, or your impression of the system while operating a machine. Please mark an "x" on each line at the point which best describes your feeling or your impression.

(Note: not at all=1; extremely=7)

1 The system is deceptive

1 2 3 4 5 6 7

2 The system behaves in an underhanded manner

1 2 3 4 5 6 7

3 I am suspicious of the system's intent, action, or outputs

1 2 3 4 5 6 7

4 I am wary of the system

1 2 3 4 5 6 7

5 The system's actions will have a harmful or injurious outcome

1 2 3 4 5 6 7

6 I am confident in the system

1 2 3 4 5 6 7

7 The system provides security

1 2 3 4 5 6 7

8 The system has integrity

1 2 3 4 5 6 7

9 The system is dependable

1 2 3 4 5 6 7

10 The system is reliable

1 2 3 4 5 6 7

11 I can trust the system

1 2 3 4 5 6 7

12 I am familiar with the system

1 2 3 4 5 6 7