AFRL-AFOSR-JP-TR-2018-0050

A decision-theoretic model of interactions between people and devices

Ingrid Zukerman
MONASH UNIVERSITY

06/15/2018
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOA
Arlington, Virginia 22203
Air Force Materiel Command

DISTRIBUTION A: Distribution approved for public release.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
15-06-2018

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*
09 Jul 2014 to 08 Feb 2018

4. TITLE AND SUBTITLE
A decision-theoretic model of interactions between people and devices

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA2386-14-1-0010

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Ingrid Zukerman

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
MONASH UNIVERSITY
WELLINGTON RD
CLAYTON, 3800 AU

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AOARD
UNIT 45002
APO AP 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-JP-TR-2018-0050

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This research aimed to obtain insights about factors that influence the acceptability of a range of devices prior to usage, i.e., users a priori attitudes towards devices. The insights gained may be employed to mitigate device disuse and design devices for specific users and tasks. The results of this research are described in (Zhan and Zukerman, 2016; Zhan et al., 2016; Zhan et al., 2018 submission pending).

The research team designed a survey that elicits users views regarding the acceptability of devices in isolation and devices in the context of tasks. The survey showed participants 17 devices, e.g., robotic cleaner, health monitoring device, humanoid robot, robotic arm, smart shirt and smart glasses; and 15 tasks potentially performed by these devices, e.g., informing of an emergency of a place or a person, adjusting the environment, performing physical actions and locating objects. The survey also obtained demographic information, information about users technical expertise and experience with devices, and interface preferences; and asked openended questions about users views regarding devices and situations where smart devices would be useful (Zhan and Zukerman, 2016). 136 people participated in the survey.

The main insights are divided into four categories: (1) devices, (2) demographic and technical experience, (3) tasks, and (4) implicit factors. The implicit factors enabled the development of recommender systems that predict users ratings of devices in isolation and devices in the context of tasks. The research team first describes their insights, followed by the results obtained by the recommender systems.

15. SUBJECT TERMS
Human Machine Trust

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON ROBERTSON, SCOTT
a. REPORT	b. ABSTRACT	c. THIS PAGE			
Unclassified	Unclassified	Unclassified	SAR		19b. TELEPHONE NUMBER *(Include area code)* +81-042-511-7008

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

DISTRIBUTION A: Distribution approved for public release.

Final Report for AOARD Grant FA2386-14-1-0010

"A decision-theoretic model of interactions between people and devices"

June 11, 2018

Name of Principal Investigators (PI and Co-PIs): Prof Ingrid Zukerman

- e-mail address : Ingrid.Zukerman@monash.edu
- Institution : Monash University
- Mailing Address : Faculty of Information Technology, Monash University,
- Clayton, Victoria 3800
- Phone : +61 3 9905 5202
- Fax : +61 3 9905 5159

Period of Performance: 06/18/2014 – 02/08/2018

Abstract

People interact with increasingly sophisticated devices that affect many aspects of their lives. For these interactions to be successful, people must bestow appropriate trust in devices. That is, people must be able to decide when they should trust a device and when they should not. This research investigates three aspects of trust:

1. **Attribute identification** – we identify attributes of people and devices that influence the initial acceptability of devices – this is an essential enabling factor, as a device must be deemed acceptable in order to be adopted.
2. **Influence of device performance –** we determine how device performance and people's attributes influence behaviour and trust in devices – once a device is adopted, trust depends to a large extent on device performance. We investigate this aspect in a complex scenario where people trade off conflicting goals.
3. **Contribution of a decision-theoretic agent –** we determine the effect of a decision-theoretic agent on people's current and learned (future) behavior and task performance, and their trust in the agent itself and in devices.

Aims, importance and achievements

1. Identification of attributes that influence acceptability

This research aims to obtain insights about factors that influence the acceptability of a range of devices prior to usage, i.e., users' *a priori* attitudes towards devices. Our insights may be employed to mitigate device disuse and design devices for specific users and tasks. The results of this research are described in (Zhan and Zukerman, 2016; Zhan *et al.*, 2016; Zhan *et al.*, 2018 – submission pending).

We designed a survey that elicits users' views regarding the acceptability of devices in isolation and devices in the context of tasks.[1] The survey showed participants 17 devices, e.g., robotic cleaner, health monitoring device, humanoid robot, robotic arm, smart shirt and smart glasses; and 15 tasks potentially performed by these devices, e.g., informing of an emergency of a place or a person, adjusting the environment, performing physical actions and locating objects. Our survey also obtained demographic information, information about users' technical expertise and experience with devices, and interface preferences; and asked open-ended questions about users' views

[1] https://monash.az1.qualtrics.com/SE/?SID=SV_8Jo8RKljwROiOJT).

regarding devices and situations where smart devices would be useful (Zhan and Zukerman, 2016). 136 people participated in our survey.

We divide our main insights into four categories: (1) devices, (2) demographic and technical experience, (3) tasks, and (4) implicit factors. The implicit factors enabled the development of recommender systems that predict users' ratings of devices in isolation and devices in the context of tasks. We first describe our insights, followed by the results obtained by the recommender systems.

1.1 Main insights

Devices

- Familiarity had a positive influence on users' initial views of devices – the health monitor and robotic cleaner had the highest ratings.
- General appearance and applicability influenced device acceptability – robotic devices were more acceptable than screen-based devices, which in turn were more acceptable than wearable devices. In addition, specific-purpose devices, e.g., robotic cleaner, were deemed more acceptable than general-purpose devices. However, when considered in the context of tasks, this distinction applied only to tasks for which the devices were designed.
- Realism was also an influential factor – realistic devices, whether physical or virtual, were the least acceptable devices.
- People's judgments regarding highly rated devices were more uniform (lower standard deviation) than their judgments of devices with lower ratings. That is, there is a high consensus about what is liked, but not about what is disliked.

Demographic attributes and technical experience

- Males, participants who were employed full time, participants with high technical expertise, and participants who owned four devices or more, respectively found devices more acceptable than females, participants who had other employment circumstances, participants with medium/low technical expertise, and participants that owned three devices or less.

Tasks

- The most generally acceptable tasks pertain to informing users about emergency conditions, while the least acceptable tasks pertain to devices performing activities autonomously.
- Devices considered in the context of tasks generally received lower ratings than devices considered in isolation. This suggests that devices that seemed acceptable on the basis of first impressions, e.g., due to familiarity or appearance, became less appealing when the participants considered them in the context of usage.

Implicit device factors

We used Principal Component Analysis (PCA) to identify latent factors that influence peoples' attitudes toward devices. These factors, listed in descending order of importance, were named as follows for devices in isolation: *clothing integrated*, *bionic physical*, *animated virtual*, *realistic humanoid*, *accessory integrated* and *functional generic*. However, we could not name the factors for devices in the context of tasks.

1.2 Recommender systems – alleviating the new-user and new-item problems

The nameable factors facilitated the development of the *PCA-LR* method, which predicts users' ratings of devices in isolation based on the direct assignment of factors to devices, and by eliciting or computing users' views about these factors. Our inability to name the factors for devices in the context of tasks prompted the development of the *KeyQ-CF* approach, which identifies informative questions whose answers facilitate the application of a Collaborative Filter (CF) to predict users' ratings of devices in the context of tasks.

We validated the latent factors, and the approaches based on them, by employing these approaches to alleviate the new-user and new-item problem in recommender systems. Specifically, we used *PCA-LR* to address both problems for devices in isolation, and *KeyQ-CF* to address the new-user problem for devices in isolation and devices in the context of tasks. The results obtained by these algorithms were compared with those obtained by two weak baselines (global and demographic recommender systems), and two strong CF-based baselines, which represent an upper bound on performance. Our main results are:

- *New devices in isolation – existing users*. *PCA-LR* outperformed the global baseline, and its performance approached that of the two strong baselines (the demographic baseline is not relevant to this problem).
- *New users – existing devices in isolation.* The performance of *PCA-LR* was better than that of the global and demographic baselines and *KeyQ-CF*, approaching the performance of the CF baselines.
- *New users – existing devices in the context of tasks.* The performance of several *KeyQ-CF* variants improved dramatically after asking a few questions, and was significantly better than that of the weak baselines. The performance of the best variant after asking only 16 questions came close to that of the strong CF baselines.

Our results show that asking a few questions about features of devices, or a few questions about devices in the context of tasks, clearly outperforms global and demographic predictions, and significantly reduces prediction error.

2. Influence of situation and device performance on trust

This research aims to understand the impact of situation and device performance on users' trust in devices. To this effect, we developed a game that takes place in a complex scenario, where people trade off rewards obtained from performing competing tasks.[2] The game is set in a retirement village, where elderly residents live in smart homes equipped with monitoring systems. Players, who "work" in the village, trade off immediate rewards received from administration tasks against potential losses incurred from attending to the residents — a setup that affords the opportunity to investigate complex behaviours. This is an important contribution of our work, which has generated significant interest at the University of Aberdeen, where PI Zukerman is spending her sabbatical. Suggested applications range from teaching ethics to persuading users to improve their cyber-security behaviour. In addition, our insights may be employed in realistic scenarios where users are trained to interact with autonomous devices, e.g., by offering a different training regime to men and women.

We obtained insights about two factors pertaining to system performance, and two user-related factors. The system-related factors are: (1) order effects of exposure to high-false-alert (highFA) and low-false-alert (lowFA) systems; and (2) influence of system performance on user behaviour, opinion about the system and self-reported trust in the system. The user-related factors are: (3) relationship between user behaviour and self-reported trust, self-perception of behaviour and task importance; and (4) influence of demographic features, technical experience and trust propensity on users' behaviour and trust in the system.

Additional contributions of this research are: the categorization of different types of players on the basis of their behavior, demographic features and trust-propensity; and the prediction of users' trust scores and aspects of behaviour from their attributes, previous behaviour and system performance.

The results of this work, obtained from 36 participants, are described (Partovi *et al.*, 2018 – submitted).

[2] http://130.56.254.85/experimentg/#.

2.1 Main insights

Insights related to system performance

- The order in which players were exposed to a monitoring system influenced their trust: when a highFA system preceded a lowFA system, users trusted the former system more than when the highFA system followed the lowFA system. This means that there is a default level of trust that is bestowed on a system, even in light of current system performance, which is modified by previous experience.
- As expected, players had a higher opinion of lowFA monitoring systems than highFA systems, and attended a higher proportion of alerts generated by lowFA monitoring systems. In addition, players had a higher level of trust in lowFA systems and in systems that didn't miss events. Interestingly, players attended significantly fewer events when they used systems that missed events than when they employed systems that didn't miss events. This means that in general, players were reactive, and didn't perform enough checks on the elderly residents. However, players tended to learn from experience, and adjusted their checking behaviour on the basis of the performance of the system they used in the previous run.

Insights related to users' behaviour

- Players were quite self-aware about their behaviour in terms of alert attendance and checks on residents, and their behaviour was fairly consistent with the importance they ascribed to the administration task and to the monitoring task. However, in terms of behaviour and trust, we only found a weak negative correlation between the proportion of attended false alerts and trust.
- Females and players with medium/high experience with smart devices had a higher level of trust in the monitoring system, attended more events, and performed more checks on residents than males and players with low experience respectively. As expected, users with a higher propensity to trust devices attended a higher proportion of alerts, but we found no correlation between trust propensity and trust scores.

We employed the K-means algorithm to cluster players according to their attributes and behaviour. The algorithm yielded three types of players: (1) most players were in the "competent" cluster, which was characterized by high earnings from the administration task, low losses from the monitoring task and high alert attendance; (2) a few players were in the "less competent" cluster, which had medium values for these parameters; and (3) a few players were in the "don't care for the elderly" cluster, which was characterized by very high monitoring losses and low alert attendance.

Our classification algorithm predicted trust scores, proportion of alerts attended, and number of checks performed on residents. We divide the most influential features into three categories: (1) user background features – trust propensity, ethnic background, age, level of education and experience with smart devices; (2) user behavioural features – trust score, proportion of alerts attended and number of false alerts attended in the previous run; and (3) system features – number of false alerts and number of missed events.

3. Contribution of a decision-theoretic agent

We devised a decision-theoretic agent to advise players in the above game. The theoretical underpinnings of the agent have been developed and implemented, and the agent has been incorporated into the game. The agent's impact on user trust and behaviour will be tested shortly (using PI Zukerman's discretionary funds). In particular, we will determine whether the agent improves users' task performance both in games where the agent is present and in subsequent games where the agent is absent; and we will compare this performance with that achieved simply through exposure to the system (without the agent).

Although this component is currently in progress, we feel that an advice-giving agent that is independent of devices, and can explain its recommendations, has significant potential. Given the large number of available devices, an independent agent that learns about the performance of a device over time can be used to augment any device, and improve the task performance of the user-device team.

List of publications and significant collaborations that resulted from this project

b) *papers published in peer-reviewed conference proceedings,*

- Zukerman, I., Partovi, A., Zhan, K., Hamacher, N., Stout, J. and Moshtaghi, M. (2017), A Game for Eliciting Trust between People and Devices under Diverse Performance Conditions. In *CGW'2017 Proceedings – IJCAI Workshop on Computer Games*. Also in *Computer Games*, T. Cazenave, M.H.M. Winands, A. Saffidine (Eds.), Communications in Computer and Information Science, Springer, 2018.
  DOI: 10.1007/978-3-319-75931-9
- Zhan, K., Zukerman, I., Moshtaghi, M. and Rees, G. (2016), Eliciting Users' Attitudes toward Smart Devices. In *UMAP2016 Proceedings – the User Modeling, Adaptation and Personalization Conference*, pp. 175-184, Halifax, Nova Scotia.
  DOI: 10.1145/2930238.2954034
- Zhan, K. and Zukerman, I. (2016), Which Smart Devices do You Like? Factors that Affect Device Acceptability. In *HAIDM2016 Proceedings – the 5th International Workshop on Human-Agent Interaction Design and Models*, New York, New York.

e) *manuscripts submitted but not yet published,*

- Partovi, A., Zukerman, I., Zhan, K., Hamacher, N. and Hohwy, J. (2018), Effect of performance-related factors on trust in automated agents in a complex scenario. Submitted to *ICMI* 2018.
- Zhan, K. and Zukerman, I. and A. Partovi (2018), Identifying and harnessing factors that influence users' attitudes toward technology. To be submitted to *User Modeling and User-Adapted Interaction*.

f) provide a list any interactions with industry or with Air Force Research Laboratory scientists or significant collaborations that resulted from this work.

- Collaborations with the University of Aberdeen are under discussion.
- Following the 2017 Program Review, I have had informal interactions with James Bliss (Old Dominion University) and Mohammad Khan (University of Connecticut) regarding trust-propensity questionnaires, which we needed for our game.

Attachments

Publications in (b) and the first manuscript in (d) have been attached; the second manuscripts in (d) will be provided shortly.