Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 3 (2015) 3070 – 3077

Procedia MANUFACTURING

6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015

Evaluation of a recommender system for single pilot operations

Arik-Quang V. Dao[a,d,]*, Kolina Koltai[b], Samantha D. Cals[b], Summer L. Brandt[a,d], Joel Lachter[a,d], Michael Matessa[c], David E. Smith[d], Vernol Battiste[a,d], Walter W. Johnson[d]

[a]*San Jose State University, San Jose, CA 95112, USA*
[b]*California State University, Northridge, Northridge, CA 91330, USA*
[c]*Rockwell-Collins, Cedar Rapids, IA 52498, USA*
[d]*NASA, NASA Ames Research Center, Moffett Field, CA 94035, USA*

Abstract

This paper discusses the quality of a recommender system implemented in a simulation to assist with choosing a diversionary airport for distressed aircraft. In the third of the series of studies investigating the feasibility of ground-supported single pilot operations (SPO) a recommender system was used by 35 airline pilots as an aid for selecting diversionary airports. These pilots, acting as ground operators, used the recommender system from a ground station when off-nominal events required them to provide ground support to a single piloted aircraft. The unique circumstances imposed by each of the scenarios required the ground operators, together with the recommender system, to consider the relative importance of different factors when recommending an airport. Post-trial questionnaires were used to evaluate the recommender system. Results indicated that the pilots did not find the recommender system very transparent and did not always trust its initial recommendation. However, pilots did appear to find the recommender system to be effective in supporting them with the high workload in off nominal situations, and interactions with the system appear to have been satisfactory. Pilots also reported in post simulation surveys a desire to have better explanations for those recommendations. Findings will inform the development of future iterations of the recommender system, as well as influence SPO procedures and further development of a prototype ground station.

Keywords: Single pilot operations; Transparency; Trust; Recommender system; Emergency landing; Ground station; Simulation

* Corresponding author. Tel.: +1-605-604-6620.
E-mail address: quang.v.dao@nasa.gov.

doi:10.1016/j.promfg.2015.07.853

1. Introduction

This paper examines operator use and evaluations of a recommender system implemented in a simulation to assist in choosing a diversionary airport for distressed aircraft. The work described here was exploratory and served primarily to inform future designs of the recommender system for implementation in reduced crew operations/single pilot operations (RCO/SPO) [1,2]. In the following sections we define the recommender system and briefly describe NASA's RCO/SPO concept, the context for which it was adapted and implemented. We conclude by reporting data collected from questionnaires designed to elicit feedback from participants about whether they agreed with the recommendations from the system. In addition, we inquired about what information influenced whether they agreed or disagreed with the recommendations and report that as well.

1.1. What is a recommender system?

A recommender system is an intelligent application that supports human information-seeking tasks by suggesting products, services, and information that best suit the needs and preferences of a user [3]. State-of-the-art recommender systems support rather than automate decision-making by employing complex algorithms that incorporate preferences, rules, and heuristics to reduce a very large variety of options into a smaller more manageable subset. This approach, as opposed to presenting a single "best" solution, allows the user to select from among that subset based on information or preferences that may not be included in the algorithm. The current authors view the development and growing popularity of recommender systems as an acknowledgement that many systems can profit by allowing humans to augment machine recommendations with quantitative and qualitative information and preferences not captured in the machine algorithm. While advanced recommender systems have found their initial home in the commercial domain, where they help people make buying decisions in the presence of a large number of alternatives, these systems have potential value in other domains. For example, in air transportation, pilots may be able to take advantage of recommenders to reduce the decision-making time for safety critical situations. However, they must be able to understand the basis for the choices being presented to them before choosing among those alternatives, as they will ultimately remain responsible for the flight. In the next section we introduce the Emergency Landing Planner (ELP), an air transportation recommender system that generates recommendations for emergency landing sites and provides explanations about the primary factors influencing the risk associated with each of the alternatives.

1.2. The Emergency Landing Planner (ELP)

The ELP is a recommender system developed at NASA to assist pilots in choosing the best emergency landing site when the aircraft is damaged or suffering from control system failures [4,5]. The ELP was initially designed for onboard use in transport category aircraft. Interface to the system was through the Flight Management System (FMS) Cockpit Display Unit (CDU), with the system being invoked through a prompt on the CDU Departure/Arrival page. Landing site recommendations were displayed on the CDU on a new page that allowed selection and examination of each alternative. Selection of an alternative caused it to become the Modified Flight Plan with the proposed route being displayed on the aircraft Navigation Display (ND). In this way, pilots could inspect several alternative recommendations before deciding on the destination and route. Once a selected route was executed, it became the Active Flight Plan in the FMS.

Landing site recommendations generated by ELP were rank ordered according to a risk value representing expected loss of life; including passenger, crew and ground casualties. We briefly mention the inputs influencing the risk calculation, but for a more extensive description of the risk model we encourage readers to see Meuleau et al. [5]. Three categories of factors served as inputs for the risk computation: (1) en route risk; (2) approach risk; (3) runway risk; and (4) airport risk. The primary factors considered for en route risk were controllability of the aircraft, distance and time to the site, complexity of the flight path, and weather along the route. Approach risk took into consideration weather along the approach path, characteristics of the instrument approach, ceiling and visibility at the airport, and population along the approach path. Runway risk included factors such as length and width of the runway, landing speed, and relative wind speed and direction. Finally, airport risk incorporated the availability of emergency facilities at and near the site. The routes generated by the ELP specified paths from the aircraft's current position to a destination runway. The following obstacles were considered in the route computations: (1) terrain; (2) hazardous weather; and (3) special use airspace.

Due to the limited screen real estate on the CDU, the ELP designers had to be selective about what information could be displayed to explain the recommendations. Explanations were limited to a short list of two character abbreviations indicating the principal risks associated with each option. For example, the code "CE" indicated that cloud ceiling was close to minimums for the approach and was a principle risk associated with the option. Had the screen real estate been larger other valuable information such as actual ceiling, visibility and winds could had been displayed for each option [4]. In the next section we describe modifications to the ELP for supporting a distressed aircraft in single pilot operations; where the ELP was hosted at a ground station that provided the needed screen area to support enhanced explanations.

2. An ELP for single pilot operations (SPO)

In an ongoing series of studies, NASA has been investigating the feasibility of single pilot operations (SPO) where the flight crew is reduced from two to one for large transport aircraft that operate under Part 121 of the Federal Airline Regulations (FARs) [1,2]. Cost-savings are the primary motivation for these operations. However, if implemented, SPO may also serve as a timely solution to a shortage in pilots predicted to occur over the next two decades [6]. The studies conducted by NASA aim to discover issues related to removing a crewmember from the flight deck, as well as investigate the efficacy of any technology used to address those issues. One approach is to re-allocate first officer roles to an operator at a ground station. There is currently no specification on whether this ground operator be a pilot, a dispatcher or some other uniquely trained personnel. However, aside from the training required, the technology supporting the ground operator role must compensate for the absence of visual cues normally used in communication when the flight crew is collocated; as well as facilitate collaboration so that the operations remain safe when the crew is separated. In light of those considerations, a ground station was developed as part of a system for evaluating SPO. Consistent with the economic justification for SPO, the ground station was designed to provide services to multiple aircraft, or remain on standby until an increase in workload triggered by off-nominal events necessitated collaboration with the flight deck. Along with a suite of other tools provided to support remote crew collaboration, participants in the simulation were asked to interact with a ground station version of the ELP to help resolve scripted off-nominal or high workload events by selecting a diversionary airport, and later provide feedback about their experience with the system.

The simulation mentioned above was the third in a series investigating SPO (SPO III). The simulation system was composed of a mid-fidelity 777 flight simulator, desktop flight simulators, controller positions, remote ground stations, and a simulation management and control position [7]. The Multi-Aircraft Control System (MACS) [8] provided the primary simulation architecture that generated and displayed scripted air traffic information. MACS also provided the flight deck and controller interfaces. Experiment confederates composed of laboratory staff members followed scripts from controller and flight deck positions to provide fidelity to the simulation. Subject matter experts (SMEs), who were experienced airline pilots, served as confederates who flew alone at flight deck stations specifically designed for open communication with experiment participants at the ground station. The SME's flight deck Mode Control Panel (MCP), CDU, ND and many of the other flight deck displays and instruments were replicated at the ground station (Figure 1a). The ground station also hosted flight tracking tools and the ELP on a display to the right of the operator (Figure 1b). Finally, tools that helped the remote crews keep track of their roles (pilot flying vs pilot not flying) and responsibilities with respects to who was handling what controls (i.e., speed, heading, altitude, or CDU) were presented to pilots on displays in area "c" of Figure 1. The principle tool of interest in this paper is the ELP. For a complete description of the tools in Figure 1a-b, see Brandt et al. [7]. For details regarding the collaboration tools in Figure 1c, see Ligda et al. [9].

The ground station version of ELP was invoked using a dedicated button located on the lower left corner of the Traffic Situation Display (TSD) in area "b" of Figure 1. When engaged in supporting one of various high workload or off nominal situations (e.g., wheel well fire), the ground operator depressed the ELP button to generate a rank-ordered list of options to discuss with the remote crew member as shown in Figure 1d. Note that an airport may

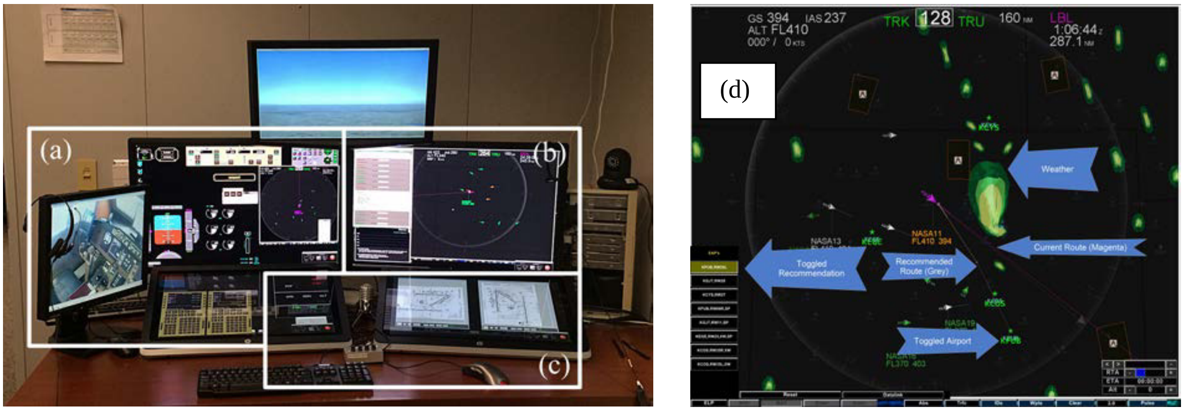

Fig. 1. The SPO III Ground Station: (a) replicated flight deck displays for the chosen aircraft; (b) flight tracking displays with ELP recommendations; and (c) crew collaboration tools including sharable charts; (d) the Traffic Situation Display and Emergency Landing Planner recommendations (lower left corner).

appear more than once in the list with different runways having different risk. Once the list of landing options was displayed the ground operator could then toggle between options to view corresponding graphical and textual explanations. The visual information and textual information highlighted the geographical location of all alternative recommendations on the TSD and graphically depicted the choices in terms of proximity to airports, complexity of the route, and constraints imposed by weather. All airports within range on the TSD scope were displayed, but dimmed out, to reduce clutter. Airports recommended by the ELP were coded in green to highlight their geographical location. Routes for a toggled option on the ELP list were rendered on the TSD as a route drawn from the nose of the aircraft symbol to the recommended airport. The current route was shown concurrent with proposed routes. When the flight deck executes a recommendation it becomes the active route and changes to magenta on the TSD. The risk model used in the ELP designed for SPO III used the same input factors used in the previous flight deck version [5].

3. Evaluation of the ELP

This evaluation of the ELP was an exploratory and informal add-on to a larger study [7] in which scripted procedures required ground station pilots to ultimately select the ELP’s top recommendation. This report draws on post-trial questionnaire data gathered from 35 participants, all professional pilots that focused on ELP issues. There were 210 responses gathered for each of these post-trial ELP related questions. Specifically, these questions were aimed at evaluating the quality of the ELP recommendations based on whether the pilots agreed with the solutions being provided by the system and whether ELP provided explanations revealed the information they needed to understand the recommendations. Tintarev and Masthoff [10] provide a set of criteria that serve to conveniently classify our evaluations. The criteria are:

- Trust;
- Transparency;
- Scrutability;
- Effectiveness;
- Persuasiveness;
- Efficiency;
- Satisfaction.

This report only addresses trust, transparency, effectiveness and efficiency. We did not address scrutability, persuasiveness, and efficiency.

3.1. Trust and transparency

Vulnerability, a common element in most definitions of trust [11–13], is particularly critical to situations of uncertainty [14]. Users are vulnerable frequently when they rely on automation; especially as automation complexity rises. The complete and timely understanding of complex automation can be impossible due to the limitations of the human perceptual and cognitive system. Even when possible, the advantages of automation would often be lost if the human must commit cognitive resources to monitoring such complexity [15]. For recommender systems it would be overwhelming to review all options considered by the system before presenting the recommended subset, and even taking the time to fully vet this subset can undo the automation's benefits. Thus, out of practicality, trust must guide reliance while much of the system remains invisible to the human [16]. Trust development requires time [14]. During that time users build trust in the system by learning under what conditions it fails and under what conditions it succeeds in accomplishing the users goals [17]. Trust is considered calibrated when it is sensitive to contexts in which the system fails or performs well [14]. The extent to which the system provides the information needed to build such trust is the system's transparency.

Tintarev [10] suggests that transparency explains how the system works. Similarly, Kim and Hinds [18] referred to transparency as understanding "why a machine behaves in an unexpected manner." Other definitions stress knowing the limits of the system through reliability information [19,20]. Chen et al. [21] defined automation transparency as "…the descriptive quality of an interface pertaining to its abilities to afford an operator comprehension about an intelligent agent's intent, performance, future plans, and reasoning processes." Two common factors that influence transparency can be inferred from these definitions: information about the system; and the human's ability to understand the information. Requiring a human to process too much information will impede the development of understanding. Too little and there is simply not enough information for transparency and the subsequent formation of trust. Evaluations of transparency for recommender systems should address whether the information actually improves understanding.

Although the scenarios required that the flight-deck confederates eventually command the execution of the top recommendation, the ground pilots, managed the ELP and transmitted its recommendations to the flight deck. Insight into whether pilots trusted the ELP to provide the best recommendation can come from how often they examined the information about other airports/runways, i.e. their acceptance of vulnerability. When asked, "How many other airports/runways did you consider?" over 80% of the pilots checked at least one other airport/runway before accepting the top choice (Figure 2a), although of these 43% only checked one airport/runway. Based on this the pilots appear to have had a fair degree of trust in the recommender system.

However, after looking at the information about other airports/runways, did pilots understand why the ELP produce the choices that it did? On the questionnaire, the pilots chose from a set of reasons (distance, runway length / width, weather, and approaches) for why the top recommendation was selected. These reasons, with the exception of "Other", mirrored the factors feeding into the ELP risk model, are shown in Figure 2b. Reasons for the top choice listed under "Other" included medical facilities, nearest suitable airport, familiarity with the airport, yielded to the captain's decision, and the nature of the emergency. In 73% of the trials pilots indicated that they understood why

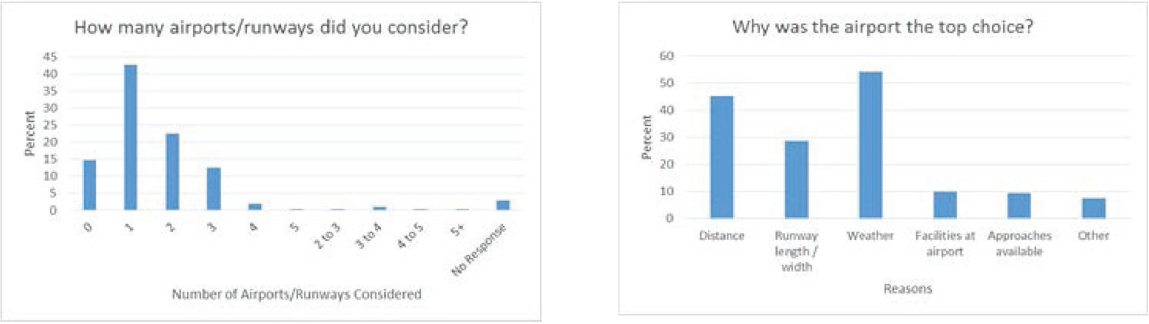

Fig. 2. (a) percent of each response category for number of airports/runways considered; (b) proportion of trials a factor was reported as a reason for the top choice.

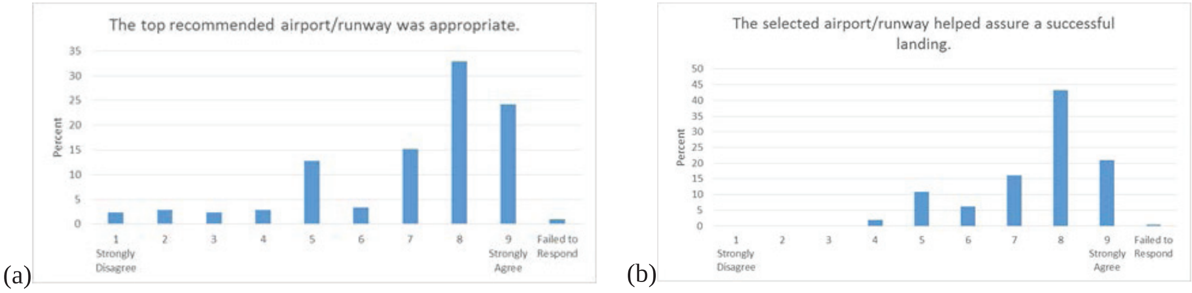

Fig. 3. (a) proportion of trials where pilots agreed with appropriateness of the top recommendation; (b) proportion of trials where pilots agreed that the top recommendation would assure a successful landing.

the top recommendation was chosen – but 27% apparently did not. It is also important that experience, not system transparency, may account for much of the pilots' ability to understand the reasons for the top recommendation.

3.2. Effectiveness

The effectiveness of the recommender system is the extent to which the system helped users make good decisions [10]. We asked three questions to acquire some insight into the effectiveness of the ELP. They were:

1. Did you deviate from the ELP provided path;
2. Was the top recommended airport/runway appropriate;
3. Did the selected airport/runway help assure a successful landing?

Responses to Question 1 showed that on 74% of the trials there were no subsequent changes to the top route recommended by the ELP. Responses to Question 2, showed that pilots mostly agreed with the appropriateness of the top recommendation (Figure 3a). Similarly, responses to Question 3 also supported effectiveness, showing that pilots were fairly confident that the top choice would result in a safe landing (Figure 3b).

3.3. Satisfaction

Satisfaction indicates whether interaction with the ELP was easy [10]. Two questions attempted to address satisfaction:

1. Would it be easy for the ground pilot to generate and choose the top recommendation without ELP assistance;
2. Did the recommender ease the burden of responding to the emergency?

Responses to Question 4 show that on 43% of the trials pilots felt they could have chosen a landing location on their own easily (Figure 4a). However, in 39% of the trials pilots failed to respond, and it is unclear if the pilots were simply unsure or if they had missed the question. However, no other question had a similar high frequency of absent responses, lending credence to the pilots not being fully confident that they could have done as well as the ELP. Supporting this, responses to Question 5 showed that, despite responses to Question 4, pilots thought that the ELP did help make it easier to respond to the overall emergency situation with responses gravitating to agree (40%) and strongly agree (20%) on most trials (Figure 4b). Here pilots failed to respond to the question on only 2% of the trials.

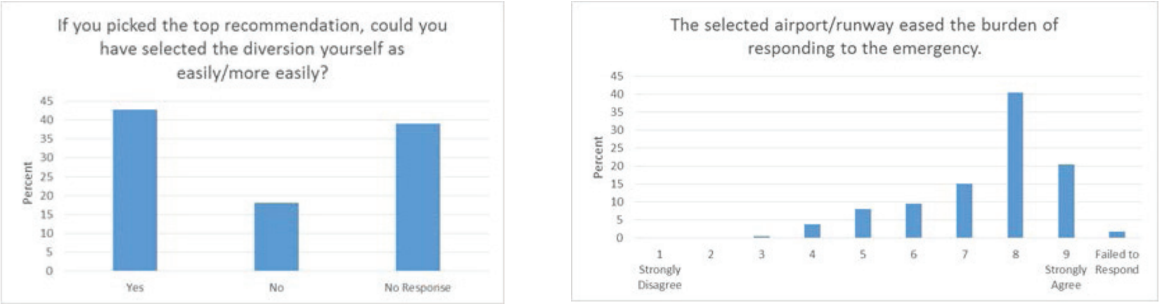

Fig. 4. (a) proportion of trials pilots felt they would have been able to pick an airport just as easily as the recommender; (b) proportion of trials pilots disagreed or agreed with whether the ELP eased overall handling of the emergency.

3.4. Post simulation debrief

At the end of each testing day we gathered pilot comments regarding their experience with ELP. Overall, pilots thought the ELP was a useful tool. We quote some of the comments that reflect that interpretation below.

- “Good tool – reduces some of the workload.”
- “Good starting point for decision-making options.”
- “Useful, once I remembered it was there.”

The remaining comments stressed that the ELP still needed to provide more explanation for its recommendations.

- “No clue why the recommendations were picked.”
- “It didn’t tell me why it didn’t take the closest airport as the nearest suitable.”
- “Transparency, did not understand the criteria for recommendations.”
- “Want to know why [the ELP] ordered the recommendation that way.”
- “Unsure of algorithm.”
- “Not familiar with criteria.”
- “Tool requires more practice to improve understanding of why it made selections.”

4. Conclusion

This paper examined the quality of the ELP recommender system as implemented in a simulation to assist ground operators with choosing a diversionary airport for distressed aircraft. Overall, the pilot feedback was positive regarding the utility of the ELP. However, in open ended debriefing discussions pilots frequently stressed that the system needed to be more transparent about how it was producing the recommendations and what inputs influenced the rank-ordering. Post-trial questions were classified according to 4 different criteria for evaluating recommender systems: (1) trust; (2) transparency; (3) effectiveness; and (4) satisfaction [10]. Only limited conclusions can be drawn from post-trial data. The fact that pilots typically only checked 1 other alternative before complying with the command to execute the top recommendation seems to point to a reasonable amount of trust in the system. However, the infrequency with which they executed the top choice without checking other alternatives might support an argument for calibrated trust in the ELP. We did not expect many instances where pilots committed to a choice without verifying against other options due to their professional training. Pilots seemed to know what information influenced the recommendations presented to them, but it was unclear if this knowledge came from experience or if that information came from system transparency. Comments made during the debriefing did make it clear that the ELP still lacked transparency; the pilots requested that the system provide more explanation for the recommendations and how the choices were ordered. Post-trial questionnaire results do suggest that the ELP was effective and pilots found it useful. Pilots favored the top recommendations generated by the ELP and in most cases did not make changes to the route once they had executed what was provided by the system. They said that the ELP did lighten the workload associated with handling the emergency events, and around 40% indicated that choosing a diversionary airport by itself was a task that they could have easily done without the ELP - but this must be weighed against the fact that about 40% of the time they did not provide an answer to that question.

Enhancements to the ELP are already under way for implementation in a follow-up simulation to demonstrate tools and concepts of operations for what will now be called reduced crew operations (RCO). In the RCO iteration the ELP (now referred to as the Autonomous Constrained Flight Planner - ACFP to reflect its new capabilities and features) will allow the user to provide new inputs. These will constrain the results according to starting state of the aircraft (e.g., position, altitude, speed etc.), the type of situation (e.g., normal, deviation, emergency type), as well as specific airports to consider, runway-length, distance to airport, time to airport, the type of approach, and altitude limitations. Operators will also be able to provide weights for a predefined set of factors: (1) en route/approach/landing/airport risk; (2) medical facilities; (3) distance; (4) fuel usage; (5) time; and (6) convenience. These weights and constraints will influence the results of the ACFP recommender and, because the operators will have provided the information, the belief is that the system will be more transparent. Finally, the ACFP will have preconfigured constraints and preferences for a defined set of situations: (1) fire, (2) medical emergency, (3) pilot incapacitation, and (4) weather diversions.

Acknowledgements

Support for this work was provided by the NASA Aviation Safety and Airspace Programs.

References

[1] J. Lachter, S.L. Brandt, V. Battiste, S. V. Ligda, M. Matessa, W.W. Johnson, in:, Proc. HCI-Aero 2014 Conf., Silicon Valley, 2014.
[2] J. Lachter, V. Battiste, M. Matessa, Q. Dao, R. Koteskey, W. Johnson, Proc. HCI-Aero 2014 Conf. (2014).
[3] T. Mahmood, F. Ricci, Proc. 20th ACM Conf. Hypertext Hypermedia - HT '09 (2009) 73.
[4] N. Meuleau, C. Neukom, C. Plaunt, D.E. Smith, T. Smith, in:, ICAPS-11 Sched. Plan. Appl. Work., 2011, pp. 1–8.
[5] N. Meuleau, C. Plaunt, D.E. Smith, in:, Proc. Twenty First Innov. Appl. Artif. Intell. Conf., AAAI Press, 2009.
[6] D. Gates, The Seattle Times (2014).
[7] S.L. Brandt, J. Lachter, V. Battiste, W.W. Johhson, in:, Pap. to Appear Proc. 6th Int. Conf. Appl. Hum. Factors Ergon., Las Vegas, NV, 2015.
[8] T. Prevot, Int. Conf. Human-Computer Interact. Aeronaut. (HCI Aero) (2002) 149.
[9] S. V. Ligda, U. Fischer, K. Mosier, M. Matessa, V. Battiste, W.W. Johnson, in:, Pap. to Appear Proc. 17th Int. Conf. Human-Computer Interact., Los Angeles, CA, 2015.
[10] N. Tintarev, J. Masthoff, in:, F. Ricci, L. Rokach, B. Shapira, P.B. Kantor (Eds.), Recomm. Syst. Handb., Springer, New York, 2011, pp. 479–507.
[11] J.L. Johns, J. Adv. Nurs. 24 (n.d.) 76.
[12] R.C. Mayer, J.H. Davis, F.D. Schoorman, 20 (2014) 709.
[13] D. Rousseau, S. Sitkin, R. Burt, C. Camerer, Acad. Manag. Rev. 23 (1998) 393.
[14] J.D. Lee, K.A. See, Hum. Factors 46 (2004) 50.
[15] C. Miller, in:, R. Shumaker, S. Lackey (Eds.), Virtual, Augment. Mix. Reality. Des. Dev. Virtual Augment. Environ. SE - 19, Springer International Publishing, 2014, pp. 191–202.
[16] J.B. Lyons, C.K. Stokes, Hum. Factors J. Hum. Factors Ergon. Soc. 54 (2011) 112.
[17] J.B. Lyons, in:, Trust Auton. Syst. Pap. from 2013 AAAI Spring Symp., 2013, pp. 48–53.
[18] T. Kim, P. Hinds, in:, Proc. 15th IEEE Int. Symp. Robot Hum. Interact. Commun., 2006, pp. 80–85.
[19] M.T. Dzindolet, S.A. Peterson, R.A. Pomranky, L.G. Pierce, H.P. Beck, Int. J. Hum.-Comput. Stud. 58 (2003) 697.
[20] L. Wang, G. A. Jamieson, J.G. Hollands, Hum. Factors J. Hum. Factors Ergon. Soc. 51 (2009) 281.
[21] J.Y.C. Chen, M. Boyce, J. Wright, K. Procci, M. Barnes, ARL Tech. Rep. (n.d.).