A Design Methodology for Trust Cue Calibration in Cognitive Agents

Ewart J. de Visser[1,2] , Marvin Cohen[1], Amos Freedy[1], and Raja Parasuraman[2]

[1] Perceptronics Solutions, 3141 Fairview Park Drive, Suite 415
Falls Church, VA 22042, USA
edevisser@percsolutions.com
[2] George Mason University, Human Factors & Applied Cognition, MSN 3F5
4400 University Drive, Fairfax VA 22030, USA

Abstract. As decision support systems have developed more advanced algorithms to support the human user, it is increasingly difficult for operators to verify and understand how the automation comes to its decision. This paper describes a design methodology to enhance operators' decision making by providing trust cues so that their *perceived* trustworthiness of a system matches its actual trustworthiness, thus yielding *calibrated trust.* These trust cues consist of visualizations to diagnose the actual trustworthiness of the system by showing the risk and uncertainty of the associated information. We present a trust cue design taxonomy that lists all possible information that can influence a trust judgment. We apply this methodology to a scenario with advanced automation that manages missions for multiple unmanned vehicles and shows specific trust cues for 5 levels of trust evidence. By focusing on both individual operator trust and the transparency of the system, our design approach allows for calibrated trust for optimal decision-making to support operators during all phases of mission execution.

Keywords: Trust, Trust Calibration, Trust Cues, Cognitive Agents, Uncertainty Visualization, Bayesian Modeling, Computational Trust Modeling, Automation, Unmanned Systems, Cyber Operations, Trustworthiness.

1 Introduction

As decision support systems have developed more advanced algorithms to support the human user, it is increasingly difficult for operators to verify and understand how the automation comes to its decision. It is therefore harder for users to diagnose the true reliability of the system and to calibrate their trust appropriately in the recommendation made by a decision-aid based on both its reliability and the mission context [1]. If users trust automation too much, they risk becoming complacent and can miss critical mistakes made by recommendations, especially when advice is unreliable [2]. Conversely, users have been shown to have high self-reliance on their own decisions even if automation performance is superior to their own [3]. Additionally, if operators trust automation too little, they may spend too much time verifying the accuracy of

R. Shumaker and S. Lackey (Eds.): VAMR 2014, Part I, LNCS 8525, pp. 251–262, 2014.

the aid, which can be costly in time-critical situations [4], [5]. Consequently, the challenge is to ensure calibrated trust in the automated aid – that is, for users to tune their trust to the aid's true capability.

Figure 1 illustrates the key issue of trust calibration. If operators are calibrated, their trust as shown on the y-axis will match the trustworthiness of the automated aid as shown on the x-axis [6]. That is, calibrated operators have trust proportional to actual automation capability, as shown by the dotted line. When operators trust automation more than it deserves, they are over-trusting the system. Operators who trust automation less than it deserves, are classified as under-trusting the system. Over-trust and under-trust are associated with misuse and disuse of automation [7].

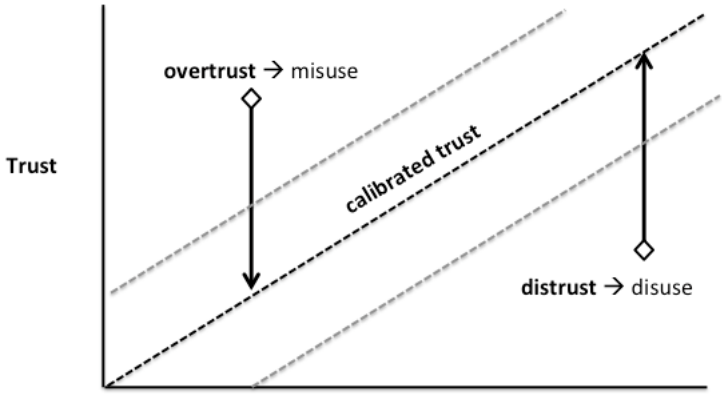


Fig. 1. Trust calibration with calibrated (green) and mis-calibrated zones (red)

We propose that providing more contextual information to the operator about the automation can lead to calibrated trust. Visual cues to diagnose the actual trustworthiness of the system by showing the risk and uncertainty of the associated information can support increased understanding of the automation. This strategy involves providing trustworthiness cues to the operator directly. By providing more information about the automation to the operator it is likely that calibration in the automation will increase [3], [6]. The effectiveness of this strategy has been shown empirically. For instance, one study showed that pilots were able to calibrate their trust in an automated aid that detected the possibility of icing conditions [8]. This automated aid showed its confidence level in its ability to estimate the likelihood of the icing condition. As a result, pilots relied on the aid more selectively and this led to improved landing performance.

The goal of this paper is to describe a trust cue design methodology to promote calibrated trust within operators. Specifically, we propose that the use of trust cues will cause over-trusting operators to reduce their trust and under-trusting operators to increase their trust into the automated system. Appropriate trust will lead to more appropriate reliance on automated systems, which will ultimately lead to better mission performance.

2 Trust Cue Design Methodology

A trust cue is any information element that can be used to make a trust assessment about an agent. In order to understand how an effective trust cue may be designed, we must first examine the trust process. We first outline five steps of a typical trust process, derive a trust cue design taxonomy based on this analysis, and describe how a trust cue may be effectively designed.

2.1 The Trust Process

The trust process can be likened to a person's attempt to construct an ad-hoc theoretical model about another human being, or generally, an agent. This process is not unlike the theory of mind abilities attributed uniquely to humans [9]. The process is updated continually over time in a loop, which can be likened to a Bayesian updating process. We have identified five steps that encompass the trust decision-making process.

Step 1: Who is to be trusted? The first step in the trust process is to consider who is to be trusted. An initial belief is formed about the object to be trusted, which has been termed dispositional-based trust [10]. During this step information is collected about the other object and is driven by both the characteristics of the trustor [10], [11], beliefs/biases about the source to be trusted. In this context, trust can be seen as a personality trait which can be measured as the propensity to trust [11], [12]. Traits of the trustor, or individual trust differences, can explain up to 52% of the variance [12]. Other biases may come from experiences, culture, etc. that have shaped general beliefs in the trustor. Knowing the source is a second great way to derive information which can, in part, explain the differences in judging machines [13], organizations [11], and interpersonal relationships [14]. Both pieces of information set the initial belief in the trustor.

Step 2: What information is available about the agent? The second consideration is what information needs to be considered to make a trust judgment. Much theorizing and classification has focused on the this step of the trust evolution, which has been labeled history-based trust [10]. The second step pertains to the evidence that is actively sampled from the trustee. As Lee and See [6] note, this can be done at various levels of abstraction using three general categories: 1) performance, which refers to the actions that are performed, 2) process, which is the consistency of those actions over time, and 3) purpose, which is the positive or negative disposition of the trustee towards the trustor. Most trust evidence can be broken down according to these dimensions. Barber [15] calls them competence, predictability, and responsibility. Mayer et al. [11] calls them ability, integrity, and benevolence. Rempel et al. [16] call them predictability, dependability, and faith. It is apparent that these different dimensions give different levels of information about the object and have different value. The order of sampling this information may occur randomly. People could start with determining the purpose, which is often the case when people are introduced to automation [6]. In such a case, one can set up their theoretical model and automatically make more assumptions then necessary which are then easily violated when observing

real behavior. In addition, information can be sampled at different levels of detail. With regard to automation, an entire system can be observed or just a sub-system. This has also been referred to as the trust resolution of a system [1].

Step 3: How to judge the available information? The third step is to evaluate and judge the presented information. This step involves calculating the belief based on the prior information and the collected information to produce a final 'posterior" trust belief, as in a Bayesian updating process. This trust belief is then evaluated (weighted) based on the constraints of the situation (completeness of information, time constraints, etc.). Lee and See [6] propose three process dimensions, including analytical, analog, and affective processes. This process can be seen as analogous to how people assess risk in general. Analytical approaches mirror the rational economic based approach. Analog processes are also called category-based trust and follow rules, expectations and contracts that have been made according to which judgments can be made. Third, the trust belief can be evaluated affectively based on affective tagging of the belief as generally favorable or unfavorable according to the risk-as-feelings, and affect heuristic paradigms [17], [18].

Step 4: What is the situation like?. The fourth step is to determine the benefit/utility of trusting the trustor compared to other alternatives (self-reliance, another person, etc.). Factors that have been found to influence this decision are self-confidence, task load, amount of effort needed, perceived risk, and exploratory behavior. This value is combined with the belief value to arrive at the utility of relying on the person to be trusted. If this value is regarded as positive, the decision will be to rely on the trustee. If it is regarded as negative, no such reliance will be observed.

Step 5: What is my reliance decision? The final step is to decide whether to rely on the automation or to seek more information. Steps one through five describe the trust decision making process. Several authors have recognized that this process is a perpetual process and repeats itself over many cycles [6], [19], [20] This has also given rise to many other classification schemes. For example, Rousseau et al. [20] propose three phases of trust including the building phase, the stability phase, and the dissolution phase. Others have proposed a dichotomous classification of trust and distrust which has been useful in classifying different sampling behaviors of evidence [19], [21]–[23]. Finally, there is an observed time phenomenon known as the inertia of trust. Lee & Moray [24] found that it took time for people to adjust their trust after experienced faults, showed residual effects over time, and found slow trust repair. In contrast to this finding, some have proposed that swift trust can be established rather quickly with a team of experts who know to look for a certain level of expertise [25].

2.2 The Trust Cue Design Taxonomy

To date there have been limited efforts to classify all information that could lead to a possible cue for trust. Typical trust dimensions are ability, integrity, and benevolence in the human-human trust literature and performance, process and purpose in the human-automation literature [6]. These bases of trust classify what task an agent is performing and how well, how the task is being performed, and the purpose/disposition

of the agent towards the trustor. These are course classifications and do not form a good basis to identify information that can lead to specific design of trust cues. There is a need to develop a rigorous taxonomy of trust cues.

Figure 2 shows our trust cue design taxonomy, with levels of trust evidence shown in the first column. With reference to the taxonomy: Intent is the overall current goal of the automation and why certain tasks are being executed. Performance is what and how well the automation is executing a particular task. Process is how an agent is executing a particular task. Expressiveness is the mode of interaction in which the automation communicates with the operator. Origin is the background information and reputation of the system. These levels of evidence include those identified in the literature as well as our own developed dimensions. In the columns are the classic information processing stages important for human performance in most tasks, including perception, comprehension, projection, decision, and execution [26]. Together these present a complete overview of the types of information that can serve as a trust cue.

Trust Cue Taxonomy		Information Processing Stages				
Trust Evidence Levels		perception	comprehension	projection	decision	execution
	Intent	perceptual intent & goals	comprehension intent & goals	projection intent & goals	decision intent & goals	execution intent & goals
	Performance	perceptual errors	classification errors	prediction errors	decision-making errors	execution errors
	Process	perceptual steps	comprehension steps	projection steps	decision steps	execution steps
	Expressiveness	perceptual indicators	comprehension indicators	projection indicators	decision indicator s	execution indicators
	Origin	design of perceptual capability	design of comprehension capability	design of projection capability	design of decision capability	design of execution capability

Fig. 2. Trust cue design taxonomy

2.3 Trust Cue Design

Given a typical trust process and our trust design taxonomy, we can now provide a procedure for designing trust cues for a specific case. The procedure is as follows:

1. Select a scenario that involves trust in a cognitive agent
2. Conduct a task analysis to identify critical trust related tasks
3. Identify key pieces of information for operator decision making
4. Verify pieces of information against the trust cue taxonomy
5. Construct a visual display representing this information

In the next section we show the results of applying this procedure to a scenario in which an automated aid assists with the management of multiple unmanned vehicles.

3 Case Study: Trust Cue Library for Unmanned Vehicle Control

3.1 Case Study: UAV Scenario

Our use case involves an On Station Operator (OSO), as a representative multi-Remotely Piloted Aircraft (RPA) operator. The OSO exercises supervisory control over all unmanned air platforms. The three main tasks of the operator include planning missions for unmanned aerial vehicles, monitoring those missions as they are executed, and reviewing these missions upon completion. Automation assistance is provided for each of these tasks and can either be fully automated or manually executed based on the level of automation [26], [27].

Since each of these tasks can be automated in our paradigm, the operator can supervise the automation performing each of these tasks. Different types of cues can be provided indicating whether the automation is succeeding in each of these tasks. In our scenario, a Situation Awareness Mixed Initiative (SAMI) module is presented as an agent acting as a subordinate mission planner that guides mission planning, execution, monitoring, and re-planning functions [28], [29]. Previous research has shown that cognitive agents can foster relationships between humans and automation [30]. Operators are first presented with a library of missions from which they can select a mission of interest. When a mission is selected, a brief description of the mission is provided to explain the actions of each UAV and the goal they will execute.

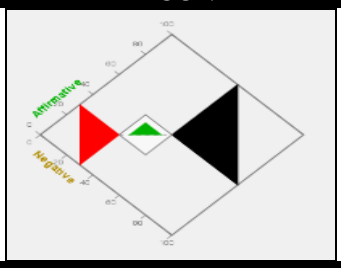
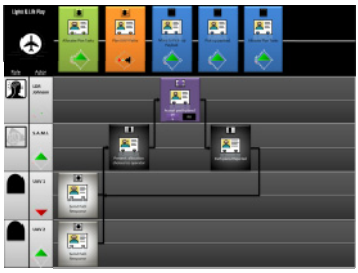
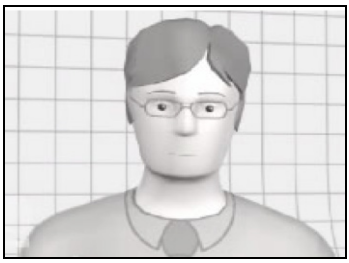
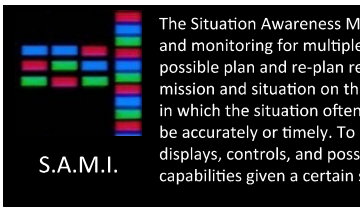
Table 1 summarizes our designs for trustworthiness cues developed according to the trust dimensions of the trust cue taxonomy and applied to the automated agent example. We elaborate on the more complex cues of performance and process in the following sections.

Table 1. Trustworthiness cue designs for each trust evidence dimension

Trustworthiness Cue	Design
Intent	**Goal Indicator**
Intent is the overall current goal of the automation and why certain tasks are being executed. The goal indicator shows the overall current goals of the agent in question. Sharing the goals of automation with a user has been shown to be an effective way to increase the trustworthiness of an adaptive cruise control system [31]; in this study, merely showing the goals of the automated agent and the user goals had the desirable effect [31].	

	S.A.M.I. Goal
	Mission success
▲	Safety
▲	Fuel efficiency
▲	Speed
▲	Accuracy

Table 1. (*Continued.*)

Performance	**DICON**
Performance is what and how well the automation is executing a particular task. The *Decision Information Icon* (DICON), a dynamic configural display, shows decomposition of uncertainty about how well the agent is performing. Showing performance feedback has been shown to increase trust and performance [8].	
Process	**Mission Model**
Process is how an agent is executing a particular task. The mission model visualization module shows how SAMI executes a model step-by-step and provides valuable insight into the complex mission models created by the SAMI automation (see Table 1). Conveying the mode of automation has been shown to increase understanding of the automation and subsequent trust [32].	
Expressiveness	**Etiquette Module**
Some have argued that one of the reasons for the trust mis-calibration with automation is that the behavior of an automated device does not conform to human-human etiquette [33], [34]. Increasing etiquette has also been shown to lead to better performance, reduced situation awareness, and improved trust [34]. Scripted avatars are a way of providing a rich user interface to communicate with operators through facial communication like gaze.	
Origin	**Reputation Module**
Origin is the background information and reputation of the system. Certificates and background information are another method to establish credibility and foster trust. This is a heuristic people often use to determine whether an agent can be trusted. The reputation module is designed to provide more background information about the automated agent. Providing a rationale and background for automation system has been shown to increase trust [3].	

3.2 Performance: DICON

Figure 3 shows the Decision Information Icon (DICON), a dynamic configural display for: (1) an "at a glance" representation of uncertainty about a hypothesis, and (2) a decomposition of the uncertainty surrounding that hypothesis into three different types based on distinctions among task input-output relationships. The DICON uses distinctive colors and locations of triangles to differentiate the performance (green), completeness of data (black), and conflict (red) components of uncertainty and uses the geometric configuration formed by the triangles to represent their current relationships. In particular, increasing incompleteness of data or team participation reduces the significance of internal variability (e.g., conflict in evidence or disagreement on the team) as well as discrimination (the preponderance of evidence or opinion in one direction or the other). These relationships are reflected in dynamics that change the size of one triangle as a function of the size of the others. Figure 3 shows how the DICON display updates over time to confirm or disconfirm the hypothesis that an automated agent is performing well.

Fig. 3. The *Decision Information Icon* (DICON) dynamic changes over time

3.3 Process: Mission Model

The mission model visualization module is comprised of several components as shown in Figure 4. These components are:

1. **Mission Phases.** The top row of the mission model shows the phases or segments of each mission. Users can click on each phase to see the detailed task structure.
2. **Mission Role Timeline Tracks.** The mission role column shows the roles for the mission and which actor fills this role. In this particular mission, there are 4 roles including the operator, the intelligent automation SAMI, and two UAVs.
3. **Decision Nodes.** Operator decisions are displayed using a special operator decision node (purple) for the mission visualization. The node can be set to either accept the automation recommendation or reject the automation recommendation.

4. **Task Status.** Petri-net transitions and arcs are shown on the timeline tracks. These elements indicate which part of the plan is currently being executed (orange), which parts have been successfully completed (green), and which parts still need to be completed (blue).

Fig. 4. The Mission Model Visualization

4 Discussion

The goal of this paper was to describe a trust cue design methodology for trust calibration in advanced automated systems. We provided background on the trust process, provided a guide for trust cue design, and gave example cases of trust cue designs.

Our focus in this paper was to describe a trust taxonomy that is domain and task independent. The trust process itself is a general process used by people on a daily basis. Most of the differences in trust is driven by the characteristics of the task and the variation in agents that assist in this task. By applying the general framework to a specific task, trust cues can be developed for any domain or task, which we believe is the strength of our approach.

The current paper described mainly visual trust cues. Our taxonomy can be applied to cover other modalities as well. Previous research has shown how best to cue humans using a cognitive-agent spectrum that varies primarily on human machine characteristics has yet to be defined [30]. These machine characteristics vary greatly

in both the real world and in studies of trust. It may thus be more useful to classify machines on a cognitive agency spectrum in which some machines are more like humans and others are closer to the traditional machine. Various trust cueing dimensions of this spectrum are visual appearance, audition, motion, and personality.

Adaptively presenting trust cues to a user could further enhance our trust calibration approach. Trust varies significantly from individual to individual. By measuring trust in real-time and adaptively showing relevant trust cues, an individual may be better able to adjust and calibrate their trust. Further research is needed to test and validate the effectiveness of this approach.

5 Conclusion

With this general trust cue design methodology, novel trust cues can be designed to assist with understanding and transparency with advanced automation for various domains. These types of designs will become increasingly important with the development of advanced automated systems. This work presents a guideline with specific examples to develop such designs.

References

1. Cohen, M., Parasuraman, R., Freeman, J.: Trust in decision aids: A model and its training implications. In: Proceedings of the 1998 Command and Control Research and Technology Symposium, pp. 1–37 (1998)
2. de Visser, E.J., Parasuraman, R.: Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. J. Cogn. Eng. Decis. Mak. 5(2), 209–231 (2011)
3. Dzindolet, M., Peterson, S., Pomranky, R.: The role of trust in automation reliance. Int. J. Hum. Comput. Stud. 58(6), 697–718 (2003)
4. Bahner, J., Huper, A., Manzey, D.: Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. Int. J. Hum. Comput. Stud. 66(9), 688–699 (2008)
5. Parasuraman, R., Manzey, D.: Complacency and Bias in Human Use of Automation: An Attentional Integration. Hum. Factors 52(3), 381–410 (2010)
6. Lee, J., See, K.: Trust in automation: Designing for appropriate reliance. Hum. Factors 46(1), 50–80 (2004)
7. Parasuraman, R., Riley, V.: Humans and automation: Use, misuse, disuse, abuse. Hum. Factors 39(2), 230–253 (1997)
8. McGuirl, J., Sarter, N.: Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. Hum. Factors 48(4), 656–665 (2006)
9. Adolphs, R.: Cognitive neuroscience of human social behaviour. Nat. Rev. Neurosci. 4, 165–178 (2003)
10. Kramer, R.: Trust and distrust in organizations: Emerging perspectives, enduring questions. Annu. Rev. Psychol. 50, 569–698 (1999)

11. Mayer, R., Davis, J., Schoorman, F.: An integrative model of organizational trust. Acad. Manag. Rev. 20(3), 709–734 (1995)
12. Merritt, S.M., Ilgen, D.R.: Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. Hum. Factors J. Hum. Factors Ergon. Soc. 50(2), 194–210 (2008)
13. Madhavan, P., Wiegmann, D.A.: Similarities and differences between human-human and human-automation trust: An integrative review. Theor. Issues Ergon. Sci. 8(4), 277–301 (2007)
14. Rotter, J.: Generalized expectancies for interpersonal trust. Am. Psychol (1971)
15. Barber, B.: The logic and limits of trust, p. 190. Rutgers University Press, New Brunswick (1983)
16. Rempel, J., Holmes, J., Zanna, M.: Trust in close relationships. J. Pers. Soc. Psychol. 49(1), 95–112 (1985)
17. Loewenstein, G., Weber, E., Hsee, C., Welch, N.: Risk as feelings. Psychol. Bull. 127(2), 267–286 (2001)
18. Slovic, P., Finucane, M., Peters, E., MacGregor, D.: The affect heuristic. Eur. J. Oper. Res. 177(3), 1333–1352 (2007)
19. Lewicki, R.J., Tomlinson, E.C., Gillespie, N.: Models of Interpersonal Trust Development: Theoretical Approaches, Empirical Evidence, and Future Directions. J. Manage. 32(6), 991–1022 (2006)
20. Rousseau, D., Sitkin, S., Burt, R.: Not so different after all: A cross-discipline view of trust. Acad. Manag. Rev. 23(3), 393–404 (1998)
21. Muir, B., Moray, N.: Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics 39(3), 429–460 (1996)
22. Jian, J., Bisantz, A., Drury, C.: Foundations for an empirically determined scale of trust in automated systems. Int. J. (2000)
23. Parasuraman, R., Sheridan, T., Wickens, C.: Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. J. Cogn. Eng. Decis. Mak. 2(2), 140–160 (2008)
24. Lee, J., Moray, N.: Trust, control strategies and allocation of function in human-machine systems. Ergonomics 35(10), 1243–1270 (1992)
25. Meyerson, D., Weick, K., Kramer, R.: Swift trust and temporary systems. In: Kramer, R., Tyler, T.R. (eds.) Trust in Organizations, pp. 166–195. Sage Publications, Inc., Thousand Oaks (1996)
26. Parasuraman, R., Sheridan, T., Wickens, C.: A model for types and levels of human interaction with automation. IEEE Trans. Syst. Man Cybern. Part A Syst. Humans 30(3), 286–297 (2000)
27. de Visser, E.J., LeGoullon, M., Freedy, A., Freedy, E., Weltman, G., Parasuraman, R.: Designing an Adaptive Automation System for Human Supervision of Unmanned Vehicles: A Bridge from Theory to Practice. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 52(4), 221–225 (2008)
28. Brooks, N., de Visser, E.J., Chabuk, T., Freedy, E., Scerri, P.: An Approach to Team Programming with Markup for Operator Interaction. In: 12th International Conference on Autonomous Agents and Multiagent Systems (2013)
29. de Visser, E.J., Kidwell, B., Payne, J., Parker, J., Chabuk, T., Scerri, P., Freedy, A.: Best of both worlds: Design and evaluation of an adaptive delegation interface. In: Human Factors and Ergonomics Society Annual Meeting (2013)

30. de Visser, E.J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., Parasuraman, R.: The World is not Enough: Trust in Cognitive Agents. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 56(1), 263–267 (2012)
31. Verberne, F.M.F., Ham, J., Midden, C.J.H.: Trust in Smart Systems: Sharing Driving Goals and Giving Information to Increase Trustworthiness and Acceptability of Smart Systems in Cars. Hum. Factors J. Hum. Factors Ergon. Soc. 54(5), 799–810 (2012)
32. Degani, A., Shafto, M., Kirlik, A.: Modes in human-machine systems: Constructs, representation, and classification. Int. J. Aviat. Psychol. 9(2), 125–138 (1999)
33. Parasuraman, R., Miller, C.A.: Trust and etiquette in high-criticality automated systems. Commun. ACM 47(4), 51–55 (2004)
34. Hayes, C., Miller, C.A.: Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology, p. 406. Auerbach Publications (2011)