
Investigating Communication Techniques to Support Trust Calibration for Automated Systems

Johann Schrammel

AIT Austrian Institute of
Technology
Vienna, Austria
johann.schrammel@ait.ac.at

Peter Fröhlich

AIT Austrian Institute of
Technology
Vienna, Austria
peter.froehlich@ait.ac.at

Alexander G. Mirnig

Center for HCI, University of
Salzburg
Salzburg, Austria
alexander.mirnig@sbg.ac.at

Olivia Dinica

AIT Austrian Institute of
Technology
Vienna, Austria
Olivia.Dinica@ait.ac.at

Andrew Lindley

AIT Austrian Institute of
Technology
Vienna, Austria
Andrew.Lindley@ait.ac.at

Robert Woitsch

BOC Asset Management
GmbH
Vienna, Austria
Robert.Woitsch@boc-eu.com

Damiano Falconi

BOC Asset Management
GmbH
Vienna, Austria
Robert.Woitsch@boc-eu.com

Matthias Baldauf

University of Applied Sciences
St. Gallen
St. Gallen, Switzerland
matthias.baldauf@fhsg.ch

Abstract

Trust in an automated system is characterized by the expectation that it will support a person in a situation characterized by uncertainty and vulnerability. It is therefore important to know in which situation one should rely on an intelligent function and when not. If the reliability of the intelligent function is underestimated or overestimated, i.e. if it is not "calibrated" well enough, it can lead to distrust or overtrust. If these phenomena occur frequently, there can be a negative impact on the long-term acceptance of intelligent applications based on advanced AI and knowledge engineering approaches. Different elements and techniques can support the calibration of trust, but their effectiveness has so far not been systematically investigated across application domains. This paper provides an overview of the state of the art on the communication of reliability, uncertainty, awareness and intent, as well as of alternatives. Furthermore, it provides first directions and an outlook into exploiting these approaches for the calibration of trust in the application area of automated driving.

Author Keywords

Trust calibration, trust, reliability uncertainty, awareness, artificial intelligence, predictive systems

CSS Concepts

• **Human-centered computing~Human computer interaction (HCI)**; *HCI theory, concepts and models*

Introduction

Through the recent advances in intelligent, AI-based technologies, close collaboration between humans and automated systems has become more widespread and effective. While this type of collaboration has many advantages, there are also several challenges such as organizing turn-taking and handover of control, addressing and act in new situations, how to express limitations in behavior—to name only a few.

A prerequisite to achieve successful cooperation is to provide humans with a solid understanding of the state and intent of the system. It is not sufficient to present the human collaborator only with the results of a computation, it is also required to have context information to understand it correctly, so that the human collaborator can adjust his expectations and levels of trust - what we call trust calibration [15].

Trust calibration is achieved when the subjective trust corresponds to the actual circumstances of the system. To achieve this, a good understanding of the elements of successful trust calibration is required. Based on our experience with the subject we consider the following elements as essential:

- the estimation of the reliability of the information (reliability)
- the estimation of associated uncertainties (uncertainty),

- understanding the system's perception and interpretation of the situation and the intended path of action (awareness & intent)
- a set of alternative scenarios that are probable or under evaluation (alternatives).

After a brief introduction into main concepts of trust and trust calibration, this paper explores and describes related work on these elements that can help to improve trust calibration, then present an overview on existing systems for trust calibration. Finally, we provide a critical discussion on open issues and future research directions.

Trust and Trust Calibration

In line with Mirnig et al [15], Ekman et al. [16] and de Visser et al [17], we conceive trust as a relation between at least two agents. This is characterized by an expectation that one or more agents (trustors) will support the achievement of another agent's (trustee) goals in a situation that is characterized by uncertainty and vulnerability.

Undertrust regarding safety of a system means that the perceived safety is lower than the actual safety. Conversely, overtrust means that the perceived safety is higher than the actual safety. According to Wagner et al. [40], these trust types can exist individually or in combination. Users can underestimate the consequences if a system fails, and/or users can underestimate the likelihood that a system will make serious mistakes at all.

Ideally, the perceived safety would be as high as the actual safety. Situations, in which neither over- nor undertrust occur, are characterized by calibrated trust.



Figure 1: Indicating Location Uncertainty, image from [19]

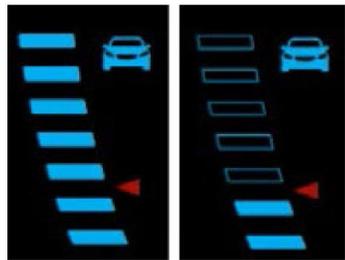


Figure 2: Reliability display on car dashboard [21]

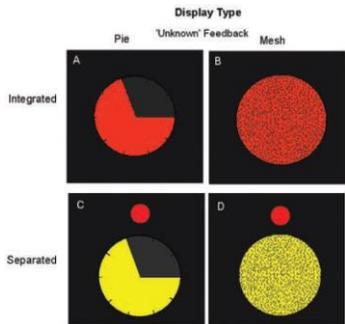


Figure 3: Indicating uncertainty in battlefield visualization [22]

In other words, trust calibration is the process of balancing user trust to the required level. If trust is not sufficiently calibrated over a longer period, users might no longer rely on the system to assist (or not sufficiently assist) them in achieving their goals in situations characterized by uncertainty and vulnerability. The next sections provide an introduction into the elements that may have a positive influence on trust calibration.

Communicating Reliability

A first important aspect of trust calibration is to directly communicate the reliability as estimated by the system to the user [26]. Typically, this consists of only one value, frequently expressed as a percentage, i.e.: "I am 75% sure the data is correct."

The display of this information needs to be tailored to the application domain, and highly different interface elements are used depending on the domain. The following examples show the wide range of possible implementations. In map visualizations, the reliability regarding location is typically shown as a circle surrounding the current position (cf figure 1). In the context of autonomous driving, the reliability of a system is shown as an indicator bar beside the main instruments on the car's dashboard (cf figure 2).

A different type of reliability display can be found in military battlefield visualization. In the example shown in figure 3, reliability of the friend/enemy detection is displayed and expressed in different ways co-located with the position information as pie-chart or color-density-coded.

Looking at these examples, the question of 2nd-order-reliability (the reliability of the reliability estimation) arises, and whether and how it should be included in the display. In the example of map visualization this would refer to the diameter of the indicated circle.

What can we learn from these examples for trust calibration? First, we think tailoring the reliability display to the specific application context is needed, and no one-size-fits-all-solution or recommendation for reliability displays can be made. Second, as can be seen in the examples, reliability is always secondary information associated with the main message, and this should be reflected in the design. Therefore, peripheral perception should be supported.

Communicating Uncertainty

Another important element of successful trust calibration is to correctly communicate the underlying level of uncertainty. Communicating uncertainty has been addressed in research in general ([10], [11]), and is a common problem in many domains, such as e.g. weather forecasts (e.g.[6],[7]) or data visualizations (e.g., [8], [9]), and learnings from these domains can be used to inform trust calibration. Figure 4 to 6 show example visualizations for communicating uncertainty in these two domains.

When communicating uncertainty, typically probabilities are used. One problem when using this approach is the problem that even well-educated adults have problems to solve easy probability questions [3]. To avoid these problems, qualitative information in labels (e.g. "low uncertainty") have been used, but they also can be misleading [4]. In addition, whether an uncertainty is

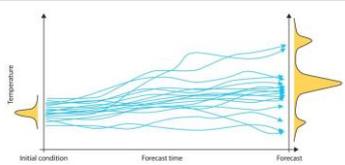


Figure 4: Showing uncertainty in ensemble weather predictions [23]

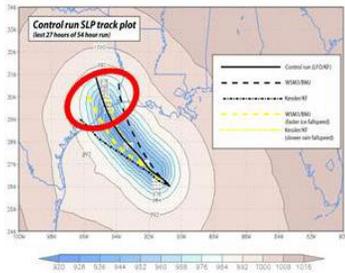


Figure 5: Visualizing in hurricane path [24]

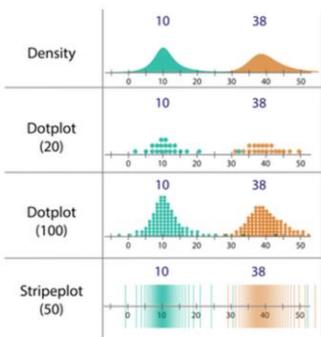


Figure 6: Different ways of showing probability distribution for uncertainty of bus departure times [25]

formulated negatively or positively also has a major influence on the following decision-making process [5].

As can be seen, no simple one-size-fits-all recommendation can be derived from prior work, and communication needs to be targeted towards the individual case. However, there is one clear finding, showing uncertainty leads to better decisions.

Communicating Awareness & Intent

Communicating awareness and intent has mainly been researched in two application domains so far: autonomous driving and human robot interaction. In autonomous driving the main focus of research is on the vehicle-pedestrian interaction, and if and how the awareness and intent of the vehicle should be communicated to the other road users.

While some studies call for explicit interfaces to communicate awareness and intent ([13]-[15]) other studies suggest that for routine situations the implicit communication (by the movement) might be sufficient [12]. Regarding trust calibration, we derive two important lessons to be learned. One must determine, first, whether it is a routine situation or not, and second, if the importance and expressivity of implicit cues (based on the observable behavior alone) are sufficient to communicate intent.

Communicating Alternatives

Another important element for successful collaboration between humans and AI systems is to communicate possible action alternatives with high probability to the human user. This allows the user to develop proper expectations regarding possible action outcomes, and

to prepare for interventions and taking over control in case the anticipated actions are problematic.

In prior work similar problems have been addressed for example from the perspective of decision support systems [2] or comparative data visualizations [1]. Regarding trust calibration, we see especially the results from the data visualization research domain as well suited to assist in designing information systems for communicating action alternatives.

Calibrating Trust

Trust calibration is to be understood in relation to over- and undertrust, as well as trust and distrust. In short, the act of calibrating trust is adjusting the user's expectations in a system such, so that neither over-, nor undertrust occurs. This means that in order to achieve this calibration, it can be necessary to induce trust or distrust in the user himself/herself at appropriate points in the interaction. Whether trust or distrust needs to be induced depends entirely on the capabilities of the system in relation to the user's expectations. If the system capabilities exceed the user's expectations, then increased trust in the system is appropriate. If, however, the inverse is the case and the system is not or insufficiently able to support the user in achieving his/her goals, then distrust is appropriate. In this regard, trust calibration requires an adjusted stance towards technologies, where both capabilities and incapacities are acknowledged and where explicitly communicating the systems incapacities is a strength rather than a weakness, as it allows the user to adjust his/her expectations, calibrate trust accordingly, and positively influence the overall interaction as a result.

While seemingly simple in principle, successfully calibrating trust in practice requires detailed knowledge of the system capabilities, interaction context, and particularly the users' prospective actions and expectations. Especially the latter can greatly vary within a context but even the system capabilities need to be appropriately specified for each level on which they intersect with the user. For the automated driving context, an initial framework by Mirnig et al. [15] proposed to break the design space down into the two dimensions of function automation (vehicle: operational, tactical, strategic) and information processing (user: perceive, understand, predict, adapt). This results in a grid of $3 \times 4 = 12$ facets to each task or maneuver in the driving context (e.g., overtaking), where for each of them the decision can be made whether the user's trust is correctly calibrated for a given situation. In case it is not, trust or distrust cues can then be targeted towards the individual facet, for a more targeted and fine-grained trust calibration process specifically for the vehicle automation context.

In 2019, Kunze et al. [19] proposed a display prototype to convey uncertainty (and thereby induced distrust) in the driver of an automated vehicle based on the principle of trust calibration. The display consisted of two primary components for the uncertainty communication: a heartbeat animation, which would change in frequency to convey the system's degree of uncertainty, together with a peripheral light strip, which would change in width and color in order to draw the driver's attention in relation to the vehicle's degree of uncertainty (from narrow to wide and blue to red to communicate increasing uncertainty and higher necessity to observe and potentially reassume control). Their results showed that safe driving performance

after a control handover was increased when using the uncertainty display, which further corroborates the hypothesis that appropriate calibration for both trust and distrust improves the interaction performance.

Conclusions and Outlook

Trust calibration can play an important role in AI-based systems to establish and guarantee their long-term acceptance. It is therefore astonishing that the design and evaluation has so far not been systematically addressed across domains. This paper has shown that different fields of research and practice has come up with a variety of techniques for communicating reliability, uncertainty, awareness and alternatives, which could eventually be used to foster trust calibration. However, a unifying approach is needed to bring together these techniques and repurpose them accordingly.

We thus further pursue the continued exploration of available design approaches and their exploitation in concrete application contexts of predictive systems in different application sectors. Under predictive systems, we subsume those that provide users with information on some historic status and that provide predictions into a future state. This can comprise predictive maintenance in industrial production, but also any kind of project monitoring and consumer systems such as in the connected home. By investigating both technical aspects of system uncertainty and user experience, we seek to obtain a holistic understanding of the topic. Based on the gathered insights, we will iteratively design and compare HCI design patterns that could be used by follow-up projects in research and industry.

Acknowledgments

This work is in part supported by the projects "auto.Bus – Seestadt" (FFG No. 860822) as part of the program "Mobilität der Zukunft" and "CALIBRaITE" FFG No. 87896) within the program "Ideenlab 4.0" that are operated by the Austrian Research Promotion Agency FFG. The financial support by the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology is gratefully acknowledged.

References

- [1] Miettinen, K. (2014). Survey of methods to visualize alternatives in multiple criteria decision making problems. *OR spectrum*, 36(1), 3-37.
- [2] Mowrer, H. T. (2000). Uncertainty in natural resource decision support systems: sources, interpretation, and importance. *Computers and electronics in agriculture*, 27(1-3), 139-154.
- [3] Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical decision making*, 21(1), 37-44.
- [4] Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, 115(4), 348.
- [5] McNeil, B. J., Pauker, S. G., Sox Jr, H. C., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England journal of medicine*, 306(21), 1259-1262.
- [6] Joslyn, S., Pak, K., Jones, D., Pyles, J., & Hunt, E. (2007). The effect of probabilistic information on threshold forecasts. *Weather and Forecasting*, 22(4), 804-812.
- [7] Morss, R. E., Demuth, J. L., & Lazo, J. K. (2019). Communicating uncertainty in weather forecasts: A survey of the US public. *Weather and forecasting*, 34(2).
- [8] Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics*, 20(12), 2142-2151.
- [9] Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016, May). When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5092-5103).
- [10] Bonneau, G. P., Hege, H. C., Johnson, C. R., Oliveira, M. M., Potter, K., Rheingans, P., & Schultz, T. (2014). Overview and state-of-the-art of uncertainty visualization. In *Scientific Visualization* (pp. 3-27). Springer, London.
- [11] Greis, M., Ohler, T., Henze, N., & Schmidt, A. (2015, September). Investigating representation alternatives for communicating uncertainty to non-experts. In *IFIP Conference on Human-Computer Interaction* (pp. 256-263). Springer, Cham.
- [12] Rothenbücher, D., Li, J., Sirkin, D., Mok, B., & Ju, W. (2016, August). Ghost driver: A field study investigating the interaction between pedestrians and driverless vehicles. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 795-802). IEEE.
- [13] Lundgren, V. M., Habibovic, A., Andersson, J., Lagström, T., Nilsson, M., Sirkka, A., ... & Saluäär, D. (2017). Will there be new communication needs when introducing automated vehicles to the urban context?. In *Advances in human aspects of transportation* (pp. 485-497). Springer, Cham.

- [14] Clamann, M., Aubert, M., & Cummings, M. L. (2017). Evaluation of vehicle-to-pedestrian communication displays for autonomous vehicles (No. 17-02119).
- [15] Alexander G. Mirnig, Philipp Wintersberger, Christine Sutter, and Jürgen Ziegler. 2016. A Framework for Analyzing and Calibrating Trust in Automated Vehicles. In Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '16 Adjunct). Association for Computing Machinery, New York, NY, USA, 33–38. DOI:https://doi.org/10.1145/3004323.3004326
- [16] F. Ekman, M. Johansson und J. L. Sochor. 2016. Creating Appropriate Trust for Autonomous Vehicle Systems: A Framework for HMI Design. Proceedings of the 95th Annual Meeting of the Transportation Research Board.
- [17] E. J. de Visser, M. Cohen, A. Freedy und R. Parasuraman. 2014. A Design Methodology for Trust Cue Calibration in Cognitive Agents. In International Conference on Virtual, Augmented and Mixed Reality.
- [18] Fröhlich, P., Schatz, R., Buchta, M., Schrammel, J., Suetterle, S., & Tscheligi, M. (2019). "What's the Robo-Driver up to?" Requirements for Screen-based Awareness and Intent Communication in Autonomous Buses. *i-com*, 18(2), 151-165.
- [19] Kunze, A., Summerskill, S. J., Marshall, R., & Filtner, A. J. (2019). Function-Specific Uncertainty Communication in Automated Driving. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 11(2), 75-97.
- [20] Yi, J., Lei, Q., Gifford, W., & Liu, J. (2017). Negative-unlabeled tensor factorization for location category inference from inaccurate mobility data. *arXiv preprint arXiv:1702.06362*.
- [21] Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications, 210-217.
- [22] Neyedli, H. F., Hollands, J. G., & Jamieson, G. A. (2011). Beyond identity: Incorporating system reliability information into an automated combat identification system. *Human factors*, 53(4), 338-355.
- [23] Grönquist, P., Ben-Nun, T., Dryden, N., Dueben, P., Lavarini, L., Li, S., & Hoefler, T. (2019). Predicting Weather Uncertainty with Deep Convnets. *arXiv preprint arXiv:1911.00630*.
- [24] Fovell, R. G. (2006). Impact of microphysics on hurricane track and intensity forecasts. In *Preprints, 7th WRF Users' Workshop, NCAR*.
- [25] Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016, May). When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5092-5103).
- [26] Noah, B. E., Gable, T. M., Chen, S. Y., Singh, S., & Walker, B. N. (2017, September). Development and preliminary evaluation of reliability displays for automated lane keeping. In Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 202-208).