Measuring and Calibrating Trust in Artificial Intelligence

Mathias Bollaert, Olivier Augereau, Gilles Coppin

Measuring and Calibrating Trust in Artificial Intelligence

Mathias Bollaert[1,2], Olivier Augereau[3,4][0000−0002−9661−3762], and Gilles Coppin[2,4][0000−0002−9193−425X]

[1] Thales DMS France
mathias.bollaert@fr.thalesgroup.com
[2] IMT-Atlantique, France
{mathias.bollaert,gilles.coppin}@imt-atlantique.fr
[3] ENIB, France
augereau@enib.fr
[4] Lab-STICC, CNRS UMR 6285, France

Abstract. Interactive systems based on Artificial Intelligence (AI) algorithms are raising new challenges, including establishing a bond of trust between users and AI. This trust must be calibrated to match the degree of reliability of AI in order to avoid over-trusting and under-trusting. However, trust is a subjective characteristic that is difficult to assess as it can vary from one person to another. This paper explores how it is possible to estimate the trust of users, especially through behavioral and physiological sensing. It also explains how, from trust assessment, it becomes possible to develop techniques for calibrating trust.

Keywords: AI · HCI · Trust · Human Factors · Trust measurement · Trust calibration

1 Introduction

Trust in Artificial Intelligence (AI) is recognised in industrial and academic circles, as a major issue [1, 19], especially due to the increasing responsibility of these systems in critical tasks such as medicine, finance, military, etc. and the potential consequences of this use, in particular in terms of safety and security, but also for ethical and legal issues.

From the end user's point of view, trust is recognized as an important element for the acceptability of systems [1] and the effectiveness of collaboration between humans and machines [7]. So, in order to match with the reliability of the system, It must be calibrated [13]. This will help to avoid over-trusting or under-trusting, both sources of dramatic accidents, especially in the history of aviation [9]. To enable this calibration, the possibility of measuring trust is an essential prerequisite.

Due to the multidisciplinary and subjective nature of trust, it is difficult to find a consensual definition of it, including in the narrower field of trust in automated systems [4], such as decision support, alert detection or automatic

classification. Still, a distinction should be made between trust which is considered as an attitude, and behaviors related to trust [19], such as reliance (user tendency to ask for a recommendation) or compliance (user tendency to follow system recommendations). Trust as an attitude [13] is inherently difficult to assess and quantify precisely, unlike behaviors that are objectively measurable, but which only have an indirect link with trust [9]. Indeed, they depend on other factors and are not necessarily reliable indicators of trust.

In the following sections of the paper we will present the most common ways to measure trust and then we will study some ways to influence and calibrate trust.

2 How to measure trust

Academic work uses different techniques to try to assess trust. We classified them into three main categories: 1) declarative, 2) behavioral, and 3) physiological.

Each category is detailed in the following subsections.

2.1 Declarative measures

A declarative measurement is subjective information collected from the users, it is also sometimes referred as attitudinal measure [18]. It can be collected through a questionnaire which can be self-reported or an interview. This information can be collected before the experiment (to assess the initial confidence), during the experiment, or after the experiment. Using standard questionnaires (such as the one proposed by Jian et al. [11]) allows a fair comparison between studies. Measurements during the experiment are often simplified in order to not interrupt the flow of the experiment. Typically, a single binary question assessing trust or distrust is asked to the users.

Another interesting measure is to ask users to assess their own self-confidence, as several studies have shown that reliance behavior is related to a relationship between user's trust and user's self-confidence [5].

2.2 Behavioral measures

Behavioral measures are quite varied in the literature. Some are generic, such as: response time, performance, error rate (and all variants, e.g. false positives / negatives), compliance [16] and reliance [12]. Other measurements are specific to the type of application. If the application involves some interactions between the user and the system, then these interactions can be measured [6]. For example it can be the time ratio during which the user delegates a task to the IA, the number of times the user uses an assistance, etc.

The most frequently used behavioral measures are response time, compliance and performance.

2.3 Physiological measures

Monitoring the user's mental state such as emotions, cognitive load, mind wandering, stress, etc. can be done by analysing physiological data [2]. Some examples of the most commonly used physiological sensors are : photoplethysmogram (PPG), electrocardiogram (ECG), electromyography (EMG), electroencephalography (EEG), electrodermal activity (EDA), etc.

If we assume that there is a correlation between the user's trust and user's mental state then, the physiological sensors might be used for estimating the user's trust.

Some recent works indicate the possibility of using an eye-tracker to assess the user's confidence [20] and trust [8, 14]. Eye Tracking is at the frontier between physiological measurement (eye blinks, pupil dilation, micro-saccades, and other uncontrolled movements) and behavioral measurement (scanpath, fixations, saccades, etc. related to visual attention and controlled movements).

Still, physiological measures are rarely used alone as they are difficult to interpret by themselves.

3 Calibrating trust

Once the trust of the user is measured, it becomes possible to try to increase or decrease it in order to calibrate it to an appropriate level. This requires the following steps:

- Assess whether one is in a case of under-trust, over-trust or appropriate trust;
- In the two former cases, determine which factors influence the level of trust and choose which ones to use;
- Use the chosen factors appropriately to adjust trust depending on the measured level and the desired level.

Determining whether we are in a situation of under-trust or over-trust may be done by comparing trust in the automation with self-confidence of the user [16]. It may also probably be directly estimated, at least in some cases, by behavioral or physiological measurements such as response times or eye fixations.

In order to influence the trust of the user, several factors have been investigated.

The main one is changing the performance or the reliability of the system, as trust is directly correlated with the performance of the system [13]. However, it is not possible to improve the reliability of the AI compared to its normal performance level, and it does not seem wise to artificially degrade it. A more usable factor which affects trust is providing reliability indicators [21]. Real-time indicators are best to provide accurate feedback to the user, but they are tricky to compute. On the other hand, global reliability indicators are easier to compute but may be less efficient in order to calibrate trust during the course of the experiment. Another factor is to use Trust Calibration Cues [16], which

alert the user when the system detects a level of trust which does not seem to be appropriate.

Bootstrapping with general reliability or explanations about how the algorithm works will have a strong impact. Indeed, if the user distrusts the system then it will be harder to rebuild the trust. In particular, in a situation of undertrust, the user would no longer use the system and therefore could not realize that it has become reliable again. Several algorithm explanations strategies has been investigated:

- Explanations on cases of false positives (for example, for a camouflaged soldier recognition app [7], the AI has trouble distinguishing a standing soldier from a natural humanoid shape, such as a tree);
- Explanations on false negatives;
- General explanations on how the algorithm works or on its strengths and weaknesses.

However, bootstrapping techniques are interesting to help providing an appropriate initial level of trust, but they do not allow dynamic trust calibration during the course of experiment. For this purpose, another possibility is to provide explanations of each results using local, post-hoc explainable AI techniques [3], such as LIME [17] (Local Interpretable Model-agnostic Explanations) or SHAP [15] (SHapley Additive exPlanations).

4 Proposal

We postulate that, to accurately measure trust, the three categories of measures must be used in conjunction. Indeed, none of these types of measurement is ideal and sufficient, which justifies combining them in order to obtain a reliable assessment of trust:

- Declarative measures involve an element of subjectivity, and extensive questionnaires are not appropriate for analyzing the evolution of trust in real-time, which is an essential component to take into account [10];
- The physiological measures are based on the hypothesis of a relationship between the level of stress and confidence, which is the subject of contradictory observations [1]. Moreover, they cannot be interpreted independently from other types of measurements. If, in an experimental environment, their use is possible, it is more difficult in real conditions, excluding perhaps eye-tracking which offers less intrusive solutions. Physiological measurement is also very difficult to interpret, generally requiring machine learning algorithms to develop predictive systems. But then, another difficulty arises from the latent nature of trust attitude: model learning is generally achieved from declarative measures, which are subjective, and thus may not accurately reflect actual trust.
- Behavioral trust-related measurements are also tricky to interpret, being influenced by many factors such as response time or fatigue. The relationship between decision time and trust, for example, is still poorly understood.

Eye Tracking, by its unique character of both behavioral and physiological measurement, and its low intrusiveness, is interesting as its measurements could be correlated with both declarative and other physiological measurements. For this reason, eye-tracking seems to us to be a particularly interesting tool for assessing trust.

We also postulate that to finely assess the evolution of trust, declarative measurements should be used during the experiment, making sure that they are the least intrusive possible. We propose to use a single and periodic analog measurement (e.g. evaluate the level of confidence between 0 and 10), in order to combine a weak intrusion and a more precise measurement than with a binary question. To our knowledge, no experiment has jointly implemented all of these criteria.

5 Conclusion and future work

Trust assessment and calibration is still an open research topic. The combination of the three categories of measurement and the implementation of measurements by eye tracking, rarely used, could make it possible to provide new means of understanding how to measure and assess trust. This prerequisite, essential to trust calibration, will then be used to experiment on the techniques of trust calibration.

We started the development of software to indirectly assess trust. The software is made of the following two parts. (1) An open-source generic instrumentation platform, that will be notably responsible for collecting and storing the various measures, providing a unified time reference, managing user profiles and experimental sessions, and for playing experimental scenarios. And (2) a specific application dedicated to the study of trust measurement. It consists of a game where the player has to rely on an AI to make decisions, and will be used to study different parameters influencing trust calibration based on the three categories of measures.

References

1. Atchley, A., Barr, H.M., O'Hear, E., Weger, K., Mesmer, B., Gholston, S., Tenhundfeld, N.: Trust in systems: Identification of 17 unresolved research questions and the highlighting of inconsistencies (2022)
2. Augereau, O., Tag, B., Kise, K.: Mental state analysis on eyewear. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. pp. 968–973 (2018)
3. Boidot, C., Augereau, O., de Loor, P., Lefort, R.: Benefits of using multiple post-hoc explanations for Machine Learning. In: 2023 International Conference on Machine Learning and Applications (ICMLA). Jacksonville, United States (2023)
4. Cohen, M.S., 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. CCRP Washington, DC (1998)

5. De Vries, P., Midden, C., Bouwhuis, D.: The effects of errors on system trust, self-confidence, and the allocation of control in route planning. International Journal of Human-Computer Studies **58**(6), 719–735 (2003)
6. Drnec, K., Marathe, A.R., Lukos, J.R., Metcalfe, J.S.: From trust in automation to decision neuroscience: applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction. Frontiers in human neuroscience **10**, 290 (2016)
7. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. International journal of human-computer studies **58**(6), 697–718 (2003)
8. Hergeth, S., Lorenz, L., Vilimek, R., Krems, J.F.: Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. Human factors **58**(3), 509–519 (2016)
9. Hoff, K.A., Bashir, M.: Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors **57**(3), 407–434 (2015)
10. Hoffman, R.R., Johnson, M., Bradshaw, J.M., Underbrink, A.: Trust in automation. IEEE Intelligent Systems **28**(1), 84–88 (2013)
11. Jian, J.Y., Bisantz, A.M., Drury, C.G.: Foundations for an empirically determined scale of trust in automated systems. International journal of cognitive ergonomics **4**(1), 53–71 (2000)
12. Lai, V., Tan, C.: On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 29–38 (2019)
13. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. Human factors **46**(1), 50–80 (2004)
14. Lu, Y., Sarter, N.: Eye tracking: a process-oriented method for inferring trust in automation as a function of priming and system reliability. IEEE Transactions on Human-Machine Systems **49**(6), 560–568 (2019)
15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
16. Okamura, K., Yamada, S.: Adaptive trust calibration for human-ai collaboration. Plos one **15**(2), e0229132 (2020)
17. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
18. Scharowski, N., Perrig, S.A., von Felten, N., Brühlmann, F.: Trust and reliance in xai–distinguishing between attitudinal and behavioral measures. arXiv preprint arXiv:2203.12318 (2022)
19. Vereschak, O., Bailly, G., Caramiaux, B.: How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW2), 1–39 (2021)
20. Yamada, K., Kise, K., Augereau, O.: Estimation of confidence based on eye gaze: an application to multiple-choice questions. In: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers. pp. 217–220 (2017)
21. Zhang, Y., Liao, Q.V., Bellamy, R.K.: Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 295–305 (2020)