Human Factors in Robots, Drones and Unmanned Systems, Vol. 138, 2024, 121–127
https://doi.org/10.54941/ahfe1005015

Human Autonomy Teaming: Proposition of a New Model of Trust

Hélène Unrein, Théodore Letouzé, Jean-Marc André, and Sylvain Hourlier

IMS – Cognitique UMR 5218, ENSC-Bordeaux INP, Talence, France

ABSTRACT

The literature regarding trust between a human and a technological system is abundant. In this context, trust does not seem to follow a simple dynamic given the multiple factors that impact it: mode of communication of the system, appearance, severity of possible system failures, factors favoring recovery, etc. In this work, we propose a modeling of the dynamics of the trust of a human agent towards an autonomous system (Human Autonomy Teaming HAT) which is inspired by a hysteresis cycle. The latter reflects a delay in the effect in the behavior of materials called inertia. According to this same principle, the variation in confidence would be based on a non-linear relationship between confidence and expectation. Thus, these variations would appear as interactions occur (like a discrete variable), rather than on a continuous time scale. Furthermore, we suggest that trust varies depending on: the conformity of expectations, the previous level of trust, the duration of maintaining a good or bad level of trust, and the interindividual characteristics of the human agent. Expectations reflect the evaluation of the situation estimated by the human agent on the basis of the knowledge at its disposal and the expected performance of the system. At each confrontation with reality, if the perceived reality agrees with the expected then the expectations are consistent, otherwise they are non-compliant. Depending on the initial state of trust, these expectations will influence the variation in trust. The latter is determined through the hysteresis cycle. At both ends of the cycle, the level of trust is characterized as either calibrated trust or distrust. Indeed, confidence does not increase towards a maximum, but towards an optimal level: calibrated confidence. This is a level of confidence adapted to the capabilities of the autonomous system. Conversely, trust decreases to a level of distrust. This corresponds to the situation where the individual does not trust the system and rejects it. In our context of use, the individual is obliged to continue to interact with the autonomous system, which opens the possibility of overcoming this distrust and restoring all or part of the initial trust. We propose that maintaining this level of calibrated trust or distrust results in an inertia effect. The more trust is maintained at one of these levels, the greater the inertia. Thus, calibrated trust established over a short period of time will be more affected by non-compliant expectations than calibrated trust established over the long term. Furthermore, the evolution of trust is influenced by individual criteria. Although the model described here is generic, it can be personalized according to the predispositions of the human agent: propensity for trust, personality trait, attitudes towards technological systems, etc. The model presented is not intended to debate the nature of trust. It illustrates and explains the dynamics of trust, a key factor in the HAT relationship, both at the origin of this interaction and for the results it produces.

Keywords: Human autonomy teaming, Trust, Human-system interaction

INTRODUCTION

Advances in artificial intelligence, robotics, automation and computer science have led to the development of more and more sophisticated autonomous systems. These technological developments are leading to an increasing hybridisation of teams (human-autonomous system), and to closer collaboration in the work process. The trust of the human agent is a prerequisite for teamwork with an autonomous system. For collaboration with a complex intelligent agent to take place, a certain level of trust must be established between the operator and the system (Saur & Ford, 1995; Zaibet, 2006).

Trust is a concept studied extensively in the literature, with a rich diversity of points of view. For example, it is sometimes associated with expectations about a person's behaviour (Deutsch, 1958), a mechanism for reducing social complexity (Luhmann, 2006), a rational choice (Orleán, 2000) or a decision to assume a risk (Mayer et al., 1995).

Several authors define types of trust that differ from one another in origin and context: situational trust, acquired trust and dispositional trust (Marsh & Dibben, 2003); assured trust and decided trust (Luhmann, 2001); calculated trust, personal trust and institutional trust (Williamson, 1993); interpersonal trust (Hardin, 2006); intuitu personae trust, relational trust and institutional trust (Zucker, 1986).

Several models explain the dimensions or dynamics of trust in human-human, human-organisation, human-machine and human-automated system interactions (Bindewald et al., 2018; Hancock et al., 2011; Lee & See, 2004; Rajaonah, 2006). Whether closed-loop or not, these models focus on the various factors that influence trust before interaction (a priori) and during interaction (a posteriori). Few authors focus their models on the causes of positive or negative variations in trust.

The model by Roxane Zolin et al. (2000) looks at the process of developing trust within organisations. The authors include the notion of evaluating performance in comparison with the operator's expectations. A positive comparison will increase trust, and vice versa. This notion of expectation is the central point of our work on variations in trust.

PROPOSITION OF TRUST MODEL

We propose a model of the dynamics of trust between a human agent and an autonomous system. We take our inspiration from the hysteresis cycle (see Figure 1) representing the state of magnetisation of a magnetic material (ordinate), according to an applied magnetic field (abscissa). In this case, the link between the state of magnetisation of the material and the magnetic field is not linear. The cycle reflects a delayed effect in the behaviour of materials called inertia: despite a reduction in the magnetic field, magnetic materials retain the memory of their previous state.

We build on this model to explain the establishment, loss, or restoration of trust, and its inertia (Desai et al., 2013) shown by the effect delays. Trust would vary according to the conformity of expectations, the previous level of trust, the duration of maintaining a good or bad level of trust, and the inter-individual characteristics of the human agent.

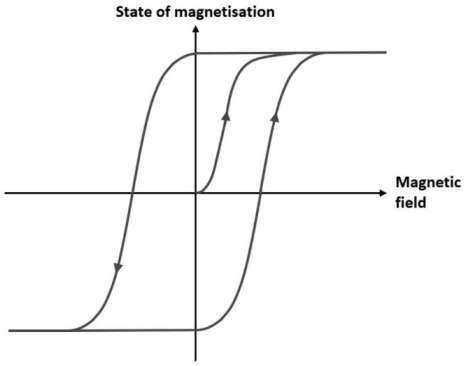

Figure 1: The hysteresis cycle (Ewing, 1882).

Conformity of Expectations

Our model describes a variation of trust based on a non-linear relationship between trust and expectation.

An operator's trust in a system is established by the regularity of the quality of execution of a specific task, mechanical understanding (the more we understand how a system works, the more comfortable we tend to be with it), predictability (anticipating the behaviour of a system increases trust in its use) and familiarity (the more effectively a system is used, the greater the trust) (Pesqueur, 2021). Trust in use therefore changes over the interaction period, based on concrete antecedents: expectations. Each time the human agent is confronted with reality, an evaluation is made between its expectations and reality: if the autonomous system meets or exceeds expectations, the expectations are said to be conformed; if the system fails to meet expectations, the expectations are said to be non-conformed. It is important to note that the term "conforms" means that what was anticipated by the individual corresponds to reality, without necessarily indicating whether the expectations are well-founded or correctly constructed on the basis of the human agent's knowledge. This other notion will be developed later.

PREVIOUS LEVEL OF TRUST

According to the literature, trust in the current moment is significantly influenced by trust in the previous moment (Lee and Moray, 1992). This is why expectations, conform or not, do not necessarily lead to a variation in trust. On the other hand, the accumulation of interactions will have an effect on trust. The variation in trust according to the conformity of expectations also depends on the previous state of trust, and is determined through the modified hysteresis cycle (see Figure 2).

Our proposition brings together the different trust variation processes identified in the literature. Once trust has been established, it is maintained, but it can be reconsidered on specific occasions (Castello, 2012; Mishra and Spreitzer, 1998). Once trust has deteriorated, it can be restored (Kim and Mauborgne, 2003).

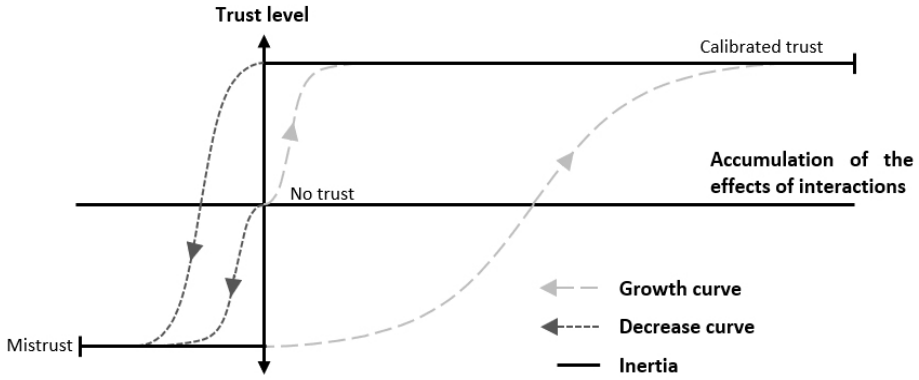

Figure 2: The modified hysteresis cycle, illustrating the variation in trust depending on the cumulative effects of the interactions and the previous state of trust.

The establishment of trust during the first interactions is represented by the growth curve starting from the origin. The acquisition of mistrust during the first interactions is represented by the decrease curve starting from the origin. Note that it is impossible to return to the origin of the axes (there is no arrow towards the point (0,0)). This is because the initial interactions have an irreversible effect: trust cannot return to its baseline state. This is consistent with the fact that experience modifies the operator's mental representation of the system: expectations, which depend on the individual's knowledge of the system, will change irreversibly.

The decline in trust is illustrated by the decrease curve. Restoration is illustrated by the growth curve. These two curves are deliberately asymmetrical: the variation is greater in the case of degradation than in the case of restoration. This asymmetry is consistent with the literature, which suggests that negative events tend to have a stronger trust-reducing effect than the trust-reinforcing effect of positive events (Guo and Yang, 2021).

INERTIA

Trust does not increase towards a maximum, but towards an optimal level, adapted to the performance of the autonomous system: calibrated trust (Wang et al., 2016). The optimal trust level correctly reflects the real skill level of the automation (Merrit and Ilgen, 2008). In this case, the trust level is adapted to the system's capabilities.

Trust decreases to a level of mistrust. The level of mistrust corresponds to the situation where the individual does not trust the other agent and rejects it because of the perception of its poor reliability. The individual reaches a maximum level of caution, or great circumspection. At this stage, the individual avoids taking decisions that expose him to potential risks. They will be disinclined, if not totally reticent, to use the system. In our context of use, the individual is obliged to continue interacting with the system, which opens up the possibility of restoring trust.

We propose that maintaining this level of calibrated trust or mistrust leads to an inertia effect. The longer trust is maintained at one of these levels, the longer the inertia. Desai (2013) finds that system errors cause a greater alteration in trust when they occur at the very beginning of the interaction than when these errors are observed after a long period of use. This inertia can be seen as resistance. Resistance to the deterioration of trust represents the case where non-conform expectations do not lead to a reduction in the level of trust and, by mirror effect of the cycle, where conform expectations do not lead to a restoration of trust.

These phases of resistance represent obstacles to the correct re-evaluation (adapted to the system's performance) of trust in the autonomous system.

INTER-INDIVIDUAL CHARACTERISTICS OF THE HUMAIN AGENT

The model described is generic, and should be customised according to the predispositions of the human agent. Some individuals are more inclined to trust than others. Dispositional trust (Marsh and Dibben, 2003) is the general tendency of an individual to trust automation. This trust is immediate and, before any use is made of it, refers to the operator's beliefs about this type of system. This dispositional confidence could modify the initial growth and decline curves of our model.

Another characteristic that we consider important is sensitivity to the expected reliability of automation. Operators with high expectations of automation reliability are more responsive to changes, whether improvements or reductions in automation reliability (Pop et al., 2015). Some individuals are therefore more responsive to change than others. This could result in fewer phases of inertia.

CONCLUSION

Other factors and individual differences, which we will not list here, contribute to trust and we presume they affect either perceived performance or expectations. The literature on trust reports a large number of studies exploring the factors that influence trust (Davis, 2019). These factors identified in the literature can be incorporated into the model by modifying the evaluation of expectation conformity. Nevertheless, the model we are proposing is not intended to unify the different representations of trust.

Furthermore, two concepts need to be integrated into this model: the notion of knowledge and that of over-trust. Expectations represent the operator's estimated projection of the evolution of the situation based on the knowledge he has, including the expected system performance. Thus, adding knowledge modifies future projections, and therefore expectations. The more knowledge the operator has of the system, the more expectations are transformed, structured, extended and optimised. Overtrust, i.e. trusting too much relative to the system's capabilities, can result from a lack of knowledge about the system (Payre, 2015). The system's imperfections allow a better calibration of the perception of the real level of reliabilty. These concepts will be incorporated into our future work.

ACKNOWLEDGMENT

The authors would like to acknowledge.

REFERENCES

Bindewald, J. M., Rusnock, C. F., & Miller, M. E. (2018). Measuring human trust behavior in human-machine teams. Advances in Human Factors in Simulation and Modeling: Proceedings of the AHFE 2017 International Conference on Human Factors in Simulation and Modeling, July 17–21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8, 47–58.

Castello, A. (2012). Trust and innovation : The role of trust in joint developments of innovative products and services [PhD Thesis]. Nice.

Davis, S. E. (2019). Individual differences in operators' trust in autonomous systems : A review of the literature. Defence Science and Technology Group (DST).

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 251–258.

Deutsch, M. (1958). Trust and suspicion. Journal of conflict resolution, 2(4), 265–279.

Ewing, J. A. (1882). On effects of retentiveness in the magnetisation of iron and steel. (preliminary notice.). Proceedings of the Royal Society of London, 34(220–223), 39–45.

Guo, Y., & Yang, X. J. (2021). Modeling and Predicting Trust Dynamics in Human–Robot Teaming : A Bayesian Inference Approach. International Journal of Social Robotics, 13(8), 1899–1909. https://doi.org/10.1007/s12369-020-00703-3

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Human factors, 53(5), 517–527.

Hardin, R. (2006). Communautés et réseaux de confiance. A. Ogien, L. Quéré (éd.), Les Moments de la confiance, Economica, 91.

Kim, W. C., & Mauborgne, R. (2003). Tipping point leadership. harvard business review, 81(4), 60–69.

Lee, J. D., & See, K. A. (2004). Trust in automation : Designing for appropriate reliance. Human factors, 46(1), 50–80.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics, 35(10), 1243–1270.

Luhmann, N. (2001). Confiance et familiarité : Problèmes et alternatives. Réseaux, 4, 15–35.

Luhmann, N. (2006). La confiance : Un mécanisme de réduction de la complexité sociale. Economica.

Marsh, S., & Dibben, M. R. (2003). The role of trust in information science and technology. Annual Review of Information Science and Technology (ARIST), 37, 465–498.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of management review, 20(3), 709–734.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal : Dispositional and history-based trust in human-automation interactions. Human Factors, 50(2), 194–210.

Mishra, A. K., & Spreitzer, G. M. (1998). Explaining how survivors respond to downsizing : The roles of trust, empowerment, justice, and work redesign. Academy of management Review, 23(3), 567–588.

Orleán, A. (2000). La théorie économique de la confiance et ses limites. La confiance en question, 59–77.
Payre, W. (2015). Conduite complètement automatisée : Acceptabilité, confiance et apprentissage de la reprise de contrôle manuel [PhD Thesis]. Paris 8.
Pesqueur, M. (2021). L'ailier de demain : Le partenariat homme-machine dans l'armée de Terre. Laboratoire de recherche sur la défense (LRD), 14.
Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual differences in the calibration of trust in automation. Human factors, 57(4), 545–556.
Rajaonah, B. (2006). Rôle de la confiance de l'opérateur dans son interaction avec une machine autonome sur la coopération humain-machine. Paris 8.
Saur, C. D., & Ford, S. M. (1995). Quality, cost-effective psychiatric treatment : A CNS—MD collaborative practice model. Archives of Psychiatric Nursing, 9(6), 332–337.
Wang, N., Pynadath, D. V., & Hill, S. G. (2016). Trust calibration within a human-robot team : Comparing automatically generated explanations. 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 109–116.
Williamson, O. E. (1993). Calculativeness, trust, and economic organization. The journal of law and economics, 36(1, Part 2), 453–486.
Zaibet, O. (2006). Collaboration dans l'entreprise et intelligence collective. 15e Conférence Internationale de Management Stratégique (AIMS).
Zolin, R., Levitt, R. E., Fruchter, R., & Hinds, P. J. (2000). Modeling & monitoring trust in virtual a/e/c teams. December, CIFE Working Paper.
Zucker, L. G. (1986). Production of trust : Institutional sources of economic structure, 1840–1920. Research in organizational behavior.