Integrating trust in automation into driver state monitoring systems

Perello-March, J, Burns, C, Elliott, M & Birrell, S

Perello-March, J, Burns, C, Elliott, M & Birrell, S 2019, Integrating trust in automation into driver state monitoring systems. in T Ahram, R Taiar, S Colson & A Choplin (eds), Human Interaction and Emerging Technologies - Proceedings of the 1st International Conference on Human Interaction and Emerging Technologies, IHIET 2019. Advances in Intelligent Systems and Computing, vol. 1018, Springer-Verlag London Ltd, pp. 344-349, 1st International Conference on Human Interaction and Emerging Technologies, Nice, France, 22/08/19.
https://doi.org/10.1007/978-3-030-25629-6_53

DOI 10.1007/978-3-030-25629-6_53
ISSN 2194-5357

Publisher: Springer

The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-030-25629-6_53

Integrating Trust in Automation into Driver State Monitoring Systems

Jaume Perello-March[1,*], Christopher Burns[1], Mark Elliot[1], Stewart Birrell[1]

[1] Warwick Manufacturing Group, The University of Warwick, Coventry, UK
{Jaume.Perello-March, C.Burns.2, M.T.Elliot, S.Birrell}@warwick.ac.uk

Abstract. Inappropriate trust in highly automated vehicles (HAVs) has been identified as one of the causes in several accidents [1]–[3]. These accidents have evidenced the need to include a Driver State Monitoring System (DSMS) [4] into those HAVs which may require occasional manual driving. DSMS make use of several psychophysiological sensors to monitor the drivers' state, and have already been included in current production vehicles to detect drowsiness, fatigue and distractions. [5]. However, DSMS have never been used to monitor Trust in Automation (TiA) states within HAVs yet. Based on recent findings, this paper proposes a new methodology to integrate TiA state-classification into DSMSs for future vehicles.

Keywords: Trust in Automation · Driver State Monitoring Systems · Highly Automated Vehicles

1 Introduction

Highly Automated Driving will change the manner in which we drive, will place new demands of their users, and as a consequence, new human factors such as TiA will arise. TiA is an attitude that leads to a behaviour - reliance - [6] from which several behavioural outcomes can result including correct use, misuse, disuse and overuse of automation [7]. In addition, TiA is a multidimensional construct [8], implying that TiA may have states. The taxonomy proposed here for DSMS would classify them as Appropriate TiA, Over-TiA and Distrust in automation.

The current standardised methodology to evaluate TiA uses self-reporting tools [9], [10]. These tools have proven to be valid and feasible but are limited to experimental scenarios and post-hoc data analysis, and do not allow active real-time measurement. Therefore, a different methodology in accordance with DSMSs capabilities needs to be developed. Recent findings in the TiA literature have suggested a promising alternative for this problem – using psychophysiology [11], [12]. Using the existing TiA scales [9], [10], this research will focus on identifying the aforementioned TiA states and their psychophysiological correlates, and investigate the use of DSMS data to train a TiA classifier using machine learning algorithms.

---

df

2 TiA Decision-Making Processes

It has been proposed that TiA is an attitude involving both affective and cognitive decision-making processes [6], [13]–[15]. Following the proposal by Lee and See [6], these can be classified as:

- Analogical: trust judgements based on rules and heuristics, making use of previously known solutions for similar current problems, other individuals' experiences, societal norms, etc. Analogical decisions are less cognitively demanding than analytic judgements, allowing faster decision-making.
- Analytical: a more accurate judgement based on active evaluations of risks and benefits of trusting an automated system. It relies on knowledge about the automated system's characteristics and performance. Analytical decision-making is more cognitively demanding and time-consuming.
- Affective: This mechanism bases trust judgements on impressions, feelings and emotions regarding the trustee. Often users may not trust automation only because they feel uneasy, lacking any kind of reasoning. Affective processes are used in high time-pressured or when cognitive resources are not available to make a judgement [6].

The use of each mechanism depends on the cognitive resources available (knowledge about the system, similar past experiences, cognitive workload, etc.) and time pressures [6]. E.g. the user may rely on analytic processing when there is sufficient time and cognitive resources available but when cognitive resources are limited and time is constrained, the user may instead rely on faster, more subconscious analogical and affective processes [15]. This model also entails that the TiA attitude generated may represent particular psychophysiological patterns. The decision-making to final behavioural outcome process from represented diagrammatically in Fig. 1, where a stimulus generates TiA decision-making leading to an attitude. Once the attitude is generated, it leads to a behaviour as the outcome of TiA, and every time this process takes place, a DSMS can learn in the loop from the outcome.

3 Modelling Trust in Automation

Previous authors have already proposed mathematical models to classify TiA [16], [17]. However these models are grounded on TiA frameworks [18], [19] which have been updated [6] and [15]. Recent developments in the Human-Robot interaction domain suggest that mathematically modelling TiA in HAVs based on mental workload and affective states is achievable [20], [21]. Therefore, it is worth considering updating and adapting these newer models to the present state-of-the-art for HAVs. The following section proposes a new methodology, building on Fig 1, which could be used to classify TiA into three states using machine-learning algorithms:

1- Appropriate trust in automation. An appropriate level of TiA means that the user's trust is calibrated accordingly with the automated system's reliability [22] – i.e. the user is aware of the system limitations and capabilities, and of the current traffic situation and uses the automated system appropriately. This is the desired state as it

leads to a correct and safe use of the HAV. Appropriate TiA means that a user will be ready to take manual control when the systems requires by being aware that the system's capabilities are limited in certain situations [6]. Another example is that the user will rely upon and activate automated driving when suggested. Under such a scenario, it can be assumed that the user will be aware of the situation, engaged in the driving task and is confident about the automated system's capabilities. This raises the possibility of classifying appropriate TiA as an expected state of relaxation (i.e. positive valence and low arousal) with mental engagement. Previous studies have successfully classified relaxation states using Support Vector Machine (SVM), Regression Tree (RT), K-Nearest Neighbour (KNN) and Bayesian Network (BN) classifiers based on a Heart Rate Variability (HRV) decrease, an increase of Heart Rate (HR), Respiration Rate (RR) and Electro-Dermal Activity (EDA) [23], [24]. This method allowed the classification of discrete emotions (anxiety, boredom, engagement, frustration and anger) with high accuracy (85.81% SVM; 83.5% RT; 75.16% KNN; and 74.03% BNT). A similar method to that used in [24], could be also implemented for EEG, EDA and eye-tracking signals to classify emotional states. Mental engagement or increased mental workload has also been successfully classified using SVM [25], [26] and Artificial Neural Networks (ANN) [27]–[29]. These classifiers were based on an increase in pre-frontal activity via ERPs from EEG, blood oxygenation increases from fNIRS, increased pupil size, fixations, saccades and reductions in blink-ratio. In addition, identification by Random Forest (RF) has been used for eye-tracking metrics [30].

2- Over-trust in automation. This occurs when TiA is not properly calibrated due to e.g. higher expectations regarding the automated system's capabilities [6]. Over-trust is probably the most undesirable and hazardous state and some of the accidents reported in the introduction are good examples [1], [2]. In this case, the user lacks knowledge regarding the actual system reliability and therefore tends to believe that the automated system will be able to appropriately contend with situations that it cannot. For example, users may not be expecting a request to take-back control and may miss or ignore this, as they are not aware of the current situation and the vehicle's performance. Therefore, over-TiA could be related to low arousal emotional states (from bored to sleepy) and a relative lack of cognitive engagement with the driving task which could be classified using SVM, RT, KNN and BN [23] based on reduced RR, EDA, HR and increased HRV. A lack of situational awareness, complacency, automation bias and even a state of drowsiness can be classified using Extreme Learning Machine (ELM) and SVM methods based on reduced pre-frontal activity via EEG or fNIRS. RF may also be used for eye-tracking parameters [30] such as increased blink ratios, reduced pupil size, reduced fixation ratio and saccadic rates [25].

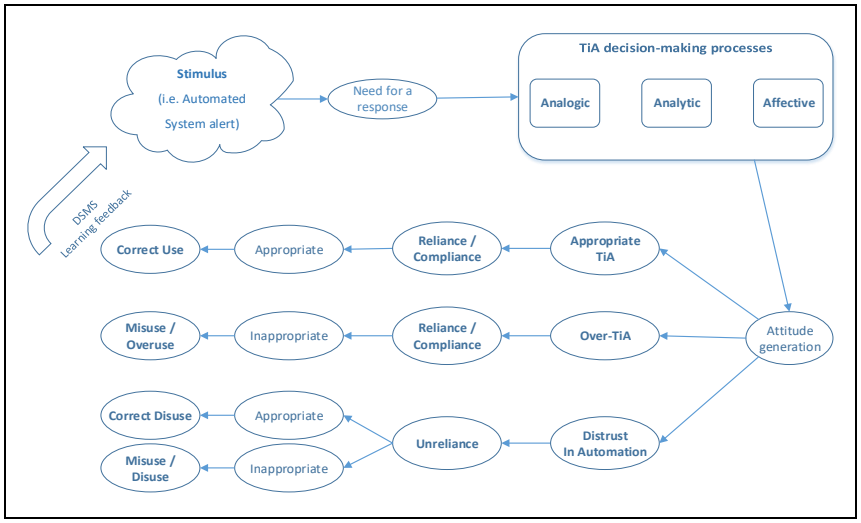

Fig. 1 Model of TiA process.

3- Distrust in Automation represents the state in which a user's self-confidence is higher than the expectations of system performance, and can be due to previous negative or inferior experiences with other, similar automated systems. Distrust may also be motivated by excessive system performance expectations that do not match the system's real capabilities, resulting in user dissatisfaction. Ultimately, the outcome of distrust usually leads to Automated System disuse [7]. Such as [3], where the HAV operator took manual control and turned right when approaching a leading vehicle, consequently colliding with an approaching motorbike on the right side. The HAV was about to reduce the speed when approaching the leading vehicle, and if the operator had not taken control, the crash would not have occurred. High-distrust situations could be associated with increased alertness or monitoring behaviours and can be classified using ELM [25], based on an increase on pre-frontal activity, pupil size, fixations, saccades and reduced blinking ratio, and RF for these visual parameters [30]. Increased mental workload can be classified using SVM [26] and ANN [27], [28], based on an increase on pre-frontal activity, increased HR, increase of tonic EDA responses, increased RR, increased pupil size, and reduced eye blinking ratio. Finally, affective states with high arousal and negative valence such as fear, distress or frustration can be expected and classified using SVM, RT, KNN and BN [23] based on increases in HR and RR, reduced HRV and increase in EDA as inputs.

Further research and conclusions

Logically, future research should focus on finding correlations between physiological data and TiA scales ([9], [10]) to identify basic TiA states (appropriate trust, distrust, and over-trust). The first study following up this paper will focus on test whether the expected psychophysiological tendencies suggested in section 3 are confirmed. Automation reliability expectations will be generated on naïve HAVs users in order to

build up different TiA among them. Using WMG's 3xD driving simulator, participants will be driven along scenarios of increasing complexity. We expect that their TiA will develop depending on the pre-administered system reliability expectations. We hypothesize that psychophysiological patterns associated with participants who trust the HAV will be statistically different from those who distrust the system. These results may serve as a TiA baseline for further related studies and as a pool of data to feed the DSMS. We aim not to find generalizable psychophysiological patterns of TiA, but to us individual patterns to train each DSMS for a particular user.

References

[1] NTSB, "Collision between a Car Operating with Automated Vehicle Control Systems and a Tractor-Semitrailer Truck. NTSB/HAR-17-XX," 2016.

[2] NTSB, "Preliminary Report HWY18MH010," p. 4, 2018.

[3] Waymo, "The very human challenge of safe driving," *Nov 5*, 2018. [Online]. Available: https://medium.com/waymo/the-very-human-challenge-of-safe-driving-58c4d2b4e8ee. [Accessed: 10-Dec-2018].

[4] V. Melnicuk, S. Birrell, E. Crundall, and P. Jennings, "Towards hybrid driver state monitoring: Review, future perspectives and the role of consumer electronics," *IEEE Intell. Veh. Symp. Proc.*, vol. 2016–Augus, pp. 1392–1397, 2016.

[5] E. Ledezma-Zavala and R. A. Ramrez-Mendoza, "Towards a new framework for advanced driver assistance systems," *Int. J. Interact. Des. Manuf.*, vol. 12, no. 1, pp. 215–223, 2018.

[6] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 46, no. 1, pp. 50–80, 2004.

[7] R. Parasuraman and V. Riley, "Humans and Automation : Use , Misuse , Disuse , Abuse," vol. 39, no. 2, pp. 230–253, 1997.

[8] R. D. Spain, E. A. Bustamante, and J. P. Bliss, "Towards an Empirically Developed Scale for System Trust: Take Two," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 52, no. 19, pp. 1335–1339, 2008.

[9] J.-Y. Jian, A. M. Bisantz, and C. Drury, "Foundations for an Empirically Determined Scale of Trust in Automated Systems," *Int. J. Cogn. Ergon.*, vol. 4, no. 1, pp. 53–71, 2000.

[10] M. Körber, "Theoretical considerations and development of a questionnaire to measure trust in automation," in *20th Triennial Congress of the IEA*, 2018, no. March, pp. 1–20.

[11] D. M. Morris, J. M. Erno, and J. J. Pilcher, "Electrodermal Response and Automation Trust during Simulated Self-Driving Car Use," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 61, no. 1, pp. 1759–1762, 2017.

[12] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *J. Exp. Soc. Psychol.*, vol. 52, pp. 113–117, 2014.

[13] E. Jan De Visser, "The World Is Not Enough: Trust in Cognitive Agents," George Mason University Fairfax, VA, 2012.

[14] M. Madsen and S. Gregor, “Measuring Human-Computer Trust,” *Proc. Elev. Australas. Conf. Inf. Syst.*, pp. 6–8, 2000.
[15] K. A. Hoff and M. Bashir, “Trust in automation: Integrating empirical evidence on factors that influence trust,” *Hum. Factors*, vol. 57, no. 3, pp. 407–434, 2015.
[16] M. Itoh and K. Tanaka, “Mathematical modeling of trust in automation: Trust, distrust, and mistrust,” *Proc. Hum. Factors Ergon. Soc. ... Annu. Meet.*, vol. 1, no. 1994, p. 9, 2000.
[17] J. Lee and N. Moray, “Trust, control strategies and allocation of function in human-machine systems,” *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992.
[18] B. M. Muir, “Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems,” *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.
[19] B. M. Muir and N. Moray, “Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation,” *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.
[20] X. Wang, Z. Shi, F. Zhang, and Y. Wang, “Mutual trust based scheduling for (semi)autonomous multi-agent systems,” *Proc. Am. Control Conf.*, vol. 2015–July, pp. 459–464, 2015.
[21] Y. Wang, Z. Shi, C. Wang, and F. Zhang, “Human-Robot Mutual Trust in (Semi)autonomous Underwater Robots,” in *Cooperative Robots and Sensor Networks*, vol. 554, no. Studies in Computational Intelligence, A. Koubaa and A. Khelil, Eds. Berlin: Springer, Berlin, Heidelberg, 2014, pp. 115–137.
[22] S. Khastgir, S. Birrell, G. Dhadyalla, and P. Jennings, “Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles,” *Transp. Res. Part C Emerg. Technol.*, vol. 96, no. July, pp. 290–303, 2018.
[23] P. Rani, C. Liu, and N. Sarkar, “An empirical study of machine learning techniques for affect recognition in human-robot interaction,” *2005 IEEE/RSJ Int. Conf. Intell. Robot. Syst. IROS*, pp. 2451–2456, 2005.
[24] J.-M. López-Gil, J. Virgili-Gomá, R. Gil, and R. García, “Corrigendum: Method for Improving EEG Based Emotion Recognition by Combining It with Synchronized Biometric and Eye Tracking Technologies in a Non-invasive and Low Cost Way,” *Front. Comput. Neurosci.*, vol. 10, no. November, pp. 9–10, 2016.
[25] L. L. Chen, Y. Zhao, J. Zhang, and J. Z. Zou, “Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning,” *Expert Syst. Appl.*, vol. 42, no. 21, pp. 7344–7355, 2015.
[26] M. A. Hogervorst, A. Brouwer, and J. B. F. Van Erp, “Combining and comparing EEG , peripheral physiology and eye-related measures for the assessment of mental workload,” vol. 8, no. October, pp. 1–14, 2014.
[27] G. F. Wilson and C. A. Russell, “Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks,” *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 45, no. 4, pp. 635–644, 2003.
[28] G. F. Wilson and C. A. Russell, “Performance Enhancement in an Uninhabited Air Vehicle Task Using Psychophysiologically Determined

Adaptive Aiding," *Hum. Factors*, vol. 49, no. 6, pp. 1005–1018, 2007.
[29] G. F. Wilson, C. A. Russell, A. Force, and A. F. Base, "Operator Functional State Classification Using Multiple Psychophysiological Features in an Air Traffic Control Task," vol. 45, no. 3, pp. 381–389, 2003.
[30] R. Zemblys, D. C. Niehorster, O. Komogortsev, and K. Holmqvist, "Using machine learning to detect events in eye-tracking data," *Behav. Res. Methods*, vol. 50, no. 1, pp. 160–181, 2018.