Handling Trust Between Drivers and Automated Vehicles for Improved Collaboration

Hebert Azevedo-Sa, X. Jessie Yang, Lionel P. Robert Jr., and Dawn M. Tilbury
University of Michigan
{azevedo,xijyang,lprobert,tilbury}@umich.edu

ABSTRACT

Advances in perception and artificial intelligence technology are expected to lead to seamless interaction between humans and robots. Trust in robots has been evolving from the theory on trust in automation, with a fundamental difference: unlike traditional automation, robots could adjust their behaviors depending on how their human counterparts appear to be trusting them or how humans appear to be trustworthy. In this extended abstract I present my research on methods for processing trust in the particular context of interactions between a driver and an automated vehicle, which has the goal of achieving higher safety and performance standards for the team formed by those human and robotic agents.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**.

KEYWORDS

Trust in Automation; Human-robot teaming; Driving simulation

ACM Reference Format:
Hebert Azevedo-Sa, X. Jessie Yang, Lionel P. Robert Jr., and Dawn M. Tilbury. 2021. Handling Trust Between Drivers and Automated Vehicles for Improved Collaboration. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3434074.3446358

1 INTRODUCTION

Trust—defined by Lee and See as "*the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability*" [16]—is a topic that has recently received considerable attention from the HRI community [17]. Trust mediates how two or more humans interact, and similarly affects how humans interact with robots or automated systems [11]. With the advance of HRI techniques, robots are expected to understand humans' behaviors that reflect trust, and adapt their own actions to better calibrate trust while they interact with those humans.

Trust in HRI has evolved from the existing large body of research on trust in the aviation and industrial processes control domains. Researchers have been trying to understand the trust-related phenomena and to model the dynamics of trust [10]. Their focus lies on: investigating the factors that influence trust; how trust evolves over time; and identifying perceivable human behaviors that reflect trust [17].

Given the high expectations regarding the development and wide use of automated vehicles (AV), the context of driver-AV interaction is particularly important, presenting research challenges with great potential for impact in safety, and AVs technology. Considering the focus on this particular HRI context, my long term research goal is to solve trust related problems that emerge when people interact with robots. To that end, my research focuses on two core questions:

- **RQ 1 -** How can an AV measure, process and influence a driver's trust (in that same AV) in order to avoid inadequate and risky reliance on the AV's capabilities, improving the driver-AV team's overall safety and performance (i.e.: how to avoid disuse and misuse of the AV, which are caused by trust miscalibration [16])?
- **RQ 2 -** Could an AV assess its own trust in the driver by observing that driver's capabilities over time? If so, can this "artificial" trust be used to better switch the driving control between the driver and the AV in different driving situations?

Therefore, my research proposes new methods that allow AVs to autonomously estimate and process both the driver's trust in the AV and the AV's trust in the human, in the hopes of ultimately improving their team-like collaboration.

2 BACKGROUND

2.1 Trust in Automation and Trust in Robots

Trust in automation is a well established topic in the fields of supervisory control and human factors engineering [14, 15, 20, 23] that has evolved from studies on interpersonal trust development [5, 21]. A common trust-related issue that potentially affects the interaction between humans and machines is *trust miscalibration* [7]. Trust miscalibration occurs when a user's trust levels are not adequate to the capabilities of the automation in use. Users can be: undertrusting the automation—when they do not use the functionalities that the machine can perform correctly because of a lack of trust; or overtrusting the automation—when, due to an excess of trust, they use the machine in situations where its capabilities are not adequate. Trust quantification [13, 20] and modeling [10, 12, 25] are especially useful for real-time detection of trust miscalibration.

As automated systems naturally evolve into robots [9], the theory on trust in automation can be extended to also characterize trust in robots. The traditional tool-operator paradigm is expected to change to a teammate-teammate paradigm, where the human and

HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8290-8/21/03.
https://doi.org/10.1145/3434074.3446358

the robot will assume the *trustor* and *trustee* positions, and vice-versa. With this transition, robots will be expected to process trust, to understand their human teammates' behaviors and to adapt their own autonomously generated actions in order for them to achieve better interaction as a team. Additionally, the teammate-teammate paradigm for HRI is likely to benefit from the establishment of robots' trust in their human counterparts. The development of trust-processing robots should fundamentally change the traditional approach for understanding and analyzing trust in automation.

2.2 Driver-AV Interaction Particularities

AVs are expected to become ubiquitous in the future as they promise to improve fuel efficiency and reduce traffic accidents. Trust in AVs is one of the main factors that influence AV adoption [3, 8]. AVs have specific characteristics that make the study of trust in driver-AV interaction challenging. For instance, people have become comfortable driving "manually" for decades, and sometimes will refuse to use self-driving capabilities—such as cruise control, automatic lane keeping or lane departure warning—because they do not understand how those capabilities work, or what are their advantages and limitations [2, 16]. Additionally, drivers usually share control with AVs, establishing a specific type of team collaboration. As in any other team, the driver and the AV must understand each other, and possibly predict intentions or control behaviors without creating vehicular instability nor increasing risks of accident.

3 MY PRIOR WORK

My previous research efforts sought to reduce the occurrence of trust miscalibration in the driver-AV interaction context. Consistent with **RQ 1**, my solution approach consisted in characterizing the driver's trust as a dynamic variable to be processed and controlled by the AV. This control engineering-based solution required the ability to estimate and to calibrate trust in real time. The lack of methods for trust estimation and calibration in the driver-AV context characterized research gaps that I have worked on to fill.

Although some work had been done in the problem of estimating human's trust in a robot [1, 18], no method was entirely appropriate for the driver-AV context. The challenge was to combine sensors to monitor the driver's behaviors that were adequate to be used in the vehicular environment with mathematical models representing the dynamics of trust over the interactions between the driver and the AV. A new estimation method was developed, which consisted of processing observable variables representing the behaviors of drivers and matching them with their self-reported levels of trust. The method relied on a Kalman filter-based solution that fused real-time data from an eye-tracking device, from the drivers' usage rate of the AV's self-driving functions, and their performance on a non-driving related task (NDRT). Experiments with a simulated SAE level 3 automated driving systems (ADS) provided the data used for model fitting. Eventually, using the estimation method, the AV was able to assess how much trust the drivers had in the AV's capabilities, based on how the drivers appeared to be splitting their attention between the driving task and the NDRT [2].

Combining the trust estimation with a trust calibration method, I have developed a trust management framework, where the AV was able to communicate with the drivers, encouraging or warning them whenever they were under- or overtrusting the AV [4]. With this framework, the AV was able to identify trust miscalibrations and to influence drivers to correct their level of trust. The trust calibrator uses the output from the trust estimator, and compares that estimate with a representation of the AV's capabilities, which changes over the specific driving contexts the AV is being operated in. Once a miscalibration is identified, the framework immediately triggers a verbal interaction from the AV to the driver, with a specific corresponding communication style and message. Those messages provide more situation awareness and a reliable risk assessment to the driver, so that they could re-calibrate their trust in the AV.

4 CURRENT AND FUTURE WORK

The literature on trust in HRI mostly addresses situations where the human is the *trustor* and the robot is the *trustee*, while little or no research considers the inverse. However, to enable seamless interactions and facilitate rapport development, both the human and the robot should be placed in both the trustor and the trustee positions. Therefore, in alignment with **RQ 2**, my research currently focuses on developing a bi-directional trust model that can represent both driver's trust in the AV and the AV's trust in the driver.

A bi-directional trust model is likely to be of fundamental importance for task allocation in human-robot teams. It mimics the team dynamics which occur when people collaborate, specifically when one teammate trusts another teammate to execute a particular set of tasks, but not others. In those situations, the trustor's trust depends on the trustee's capabilities. In the bi-directional trust model, those capabilities should represent the requirements for the execution of a task. For instance: what are the required cognitive, physical and sensory capabilities for driving on a well signalized straight road? And how do those requirements change for an unsignalized dirt road in bad weather conditions? In a more general situation, those requirements may even include non-performance factors that are known to affect trust (e.g.: trustee's characteristics [19, 22]).

Defining metric spaces to represent a task by γ, and the agent a's capabilities by λ^a, the outcome Ω of the execution of γ by a can be a success ($\Omega = 1$) or a failure ($\Omega = 0$). Considering the uncertainty involved in λ^a, trust can be computed as the probability of success when the agent a is to execute the task γ [6], denoted by

$$\tau_{\gamma,a} = P(\Omega = 1|\gamma, a) = \int_{\Lambda^a} p(\Omega = 1|\gamma, \lambda^a) bel(\lambda^a) d\lambda. \qquad (1)$$

My current work is to define Bayesian processes for dynamically updating the belief $bel(\lambda^a)$ [24] and functions that properly represent $p(\Omega|\gamma, \lambda^a)$. This model could be used not only for the driver-AV context, but also for other classes of human-robot interactions.

In my future efforts, I will use this trust model for assigning tasks for members of human-robot teams, eventually enabling negotiations between humans and robots. If a robot does not trust the human to execute a task, it must at least be able to explain its reasoning. While this negotiation strategy may raise discussions about the robots' authority, whether or not it can improve human-robot collaboration deserves a thorough investigation. Teamwork is certainly improved when team members are assigned tasks adequate to their capabilities and when other team-members trust them to execute those tasks.

REFERENCES

[1] Kumar Akash, Wan-Lin Hu, Neera Jain, and Tahira Reid. 2018. A Classification Model for Sensing Human Trust in Machines Using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems* 8, 4 (nov 2018), 1–20. https://doi.org/10.1145/3132743

[2] Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, Connor T Esterwood, X Jessie Yang, Lionel P Robert, and Dawn M Tilbury. 2020. Real-time estimation of drivers' trust in automated driving systems. *International Journal of Social Robotics* (2020), 1–17.

[3] Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, Connor T Esterwood, X Jessie Yang, Lionel P Robert Jr, and Dawn M Tilbury. 2020. Comparing the Effects of False Alarms and Misses on Humans' Trust in (Semi) Autonomous Vehicles. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 113–115.

[4] Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, X Jessie Yang, Lionel P Robert, and Dawn M Tilbury. 2020. Context-Adaptive Management of Drivers' Trust in Automated Vehicles. *IEEE Robotics and Automation Letters* 5, 4 (2020), 6908–6915.

[5] Bernard Barber. 1983. *The logic and limits of trust*. Vol. 96. Rutgers University Press, New Brunswick, NJ.

[6] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2018. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Chicago, IL, 307–315.

[7] Anders BH Christensen, Christian R Dam, Corentin Rasle, Jacob E Bauer, Ramlo A Mohamed, and Lars Christian Jensen. 2019. Reducing Overtrust in Failing Robotic Systems. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 542–543.

[8] Na Du, Jacob Haspiel, Qiaoning Zhang, Dawn Tilbury, Anuj K Pradhan, X Jessie Yang, and Lionel P Robert Jr. 2019. Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation research part C: emerging technologies* 104 (2019), 428–442.

[9] Peter A Hancock. 2017. Imposing limits on autonomous systems. *Ergonomics* 60, 2 (2017), 284–291.

[10] Kevin Hoff and Masooda Bashir. 2013. A theoretical model for trust in automated systems. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*. ACM Press, New York, New York, USA, 115. https://doi.org/10.1145/2468356.2468378

[11] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[12] Wan-Lin Hu, Kumar Akash, Tahira Reid, and Neera Jain. 2018. Computational Modeling of the Dynamics of Human Trust During Human-Machine Interactions. *IEEE Transactions on Human-Machine Systems* 1, 1 (2018), 1–13. https://doi.org/10.1109/THMS.2018.2874188

[13] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (mar 2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[14] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (oct 1992), 1243–1270. https://doi.org/10.1080/00140139208967392

[15] John D. Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies* 40, 1 (jan 1994), 153–184. https://doi.org/10.1006/IJHC.1994.1007

[16] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (jan 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[17] Michael Lewis, Katia Sycara, and Phillip Walker. 2018. The role of trust in human-robot interaction. In *Foundations of trusted autonomy*. Springer, Cham, 135–159.

[18] Yidu Lu and Nadine Sarter. 2019. 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 (2019), 560–568.

[19] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[20] Bonnie M. Muir and Neville Moray. 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 3 (mar 1996), 429–460. https://doi.org/10.1080/00140139608964474

[21] John K Rempel, John G Holmes, and Mark P Zanna. 1985. *Trust in Close Relationships*. Technical Report 1. APA. 95–112 pages. https://pdfs.semanticscholar.org/4727/fcf320e6f8c3a8bbd9d7bac22708825f48ad.pdf

[22] F David Schoorman, Roger C Mayer, and James H Davis. 2007. An integrative model of organizational trust: Past, present, and future.

[23] T.B. Sheridan, T. Vámos, and S. Aida. 1983. Adapting automation to man, culture and society. *Automatica* 19, 6 (nov 1983), 605–612. https://doi.org/10.1016/0005-1098(83)90024-9

[24] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. 2005. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press.

[25] Daniel Ullman and Bertram F Malle. 2018. What does it mean to trust a robot?: Steps toward a multidimensional measure of trust. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 263–264.