THEORETICAL FRAMEWORK FOR ANALYSIS AND EVALUATION OF HUMAN'S OVERTRUST IN AND OVERRELIANCE ON ADVANCED DRIVER ASSISTANCE SYSTEMS

Toshiyuki Inagaki and Makoto Itoh
Department of Risk Engineering, University of Tsukuba
Tsukuba 305-8573 JAPAN
inagaki@risk.tsukuba.ac.jp, itoh@risk.tsukuba.ac.jp

ABSTRACT: This paper gives a theoretical framework to describe, analyze, and evaluate the driver's overtrust in and overreliance on ADAS. Although 'overtrust' and 'overreliance' are often used as if they are synonyms, this paper differentiate the two notions rigorously. To this end, two aspects, (1) situation diagnostic aspect and (2) action selection aspect, are introduced. The first aspect is to describe overtrust, and has three axes: (1-1) dimension of trust, (1-2) target object, and (1-3) chances of observation. The second aspect, (2), is to describe overreliance on the ADAS, and distinguishes the following two types of decisions: (2-1) commission like action selection decision and (2-2) omission like action selection decision.

1   INTRODUCTION

Driving a car requires a continuous process of perception, cognition, action selection, and action implementation. Various functions are implemented in an Advanced Driver Assistance System (ADAS) to assist a human to drive a car in a dynamic environment. Such functions, sometimes arranged in a multi-layered manner, include: (a) perception enhancement that helps the driver to perceive the traffic environment around his/her vehicle, (b) arousing attention of the driver to encourage paying attention to potential risks around his/her vehicle, (c) setting off a warning to encourage the driver to take a specific action to avoid an incident or accident, and (d) automatic safety control that is activated when the driver takes no action even after being warned or when the driver's control action seems to be insufficient [1]. The first two functions, (a) and (b), are to help the driver to understand the situation. Understanding of the current situation determines what action needs to be done [2]. Once situation diagnostic decision is made, action selection decision is usually straightforward, as has been suggested by recognition-primed decision making research [3]. However, the driver may sometimes feel difficulty in action selection decision. Function (c) is to help the driver in such a circumstance. Note that any ADAS that uses only the three functions, (a) – (c), is completely compatible with the human-centered automation principle [4] in which the human is assumed to have the final authority over the automation.

Suppose an ADAS contains the forth function, (d). Then the ADAS may not always be fully compatible with the human-centered automation principle, because the system can implement an action that is not ordered by the driver explicitly. Some automatic safety control functions have been already implemented in the real world. Typical examples are seen in a pre-crash safety system (PCS) and a lane departure prevention system (LDP). PCS tightens the

seat belt and adds a warning to urge the driver to put on the brake. When the system determines that the driver is late in braking, it applies the brake automatically based on its decision. LDP is an automatic system that applies the brakes to individual wheels, without any intervention of the driver, to prevent the vehicle from departing the lane. The fact that the driver may not always be kept as the final authority over the automation in such ADAS does not necessarily mean that those designs should be prohibited. On the contrary, the automatic safety control functions are effective and indispensable to attain driver safety, which suggests the domain-dependence of human-centered automation [5]. It is true, however, that careful investigations are needed regarding to what extent the system may be given authority for deciding and act autonomously without asking the human driver's approval or consent, because autonomy of smart machines sometimes bring negative effects, such as the out-of-the-loop performance problem, loss of situational awareness, complacency or overtrust, automation surprises; see, e.g., [7-11].

Moreover, as for the forth function, (d), the following question is frequently asked: "When the ADAS is capable of coping with the situation automatically without any intervention of a driver, is not it possible for the driver to be overly reliant on the system and give up active involvement in driving?" For instance, the Ministry of Land, Infrastructure and Transport as well as the National Police Agency of the Government of Japan have been somewhat discreet in introducing highly automatic safety control functions into ADAS on concern that the drivers may place 'overtrust' in or 'overreliance' on automation. However, discussions regarding overtrust and overreliance have not been rigorous enough yet until this point. As ADAS becomes smarter and more autonomous, these issues attract more serious concerns world-wide; e.g., ASV project in Japan, HAVEit and ISi-PADAS projects in EU.

Aviation domain has various studies regarding overreliance on automation; see, e.g., [12-15]. Suppose that the automation is very rare to miss detections (i.e., it almost always alerts the human when an anomaly or an undesirable event occurs). Although a given alert is likely to be false, the human can be confident that there is no undesirable event as long as no alert is given. The human accordingly does not take precautions while the automation gives no alert. Meyer [14] has used the term reliance to express such a response of the human. If the human assumed that the automation will always give alerts when an undesired event occurs, that may be overtrust in the automation's capabilities, and the resulting reliance on the automation can be overreliance.

The relevant term, complacency, is usually defined as "self-satisfaction especially when accompanied by unawareness of actual dangers or deficiencies" [16]. However, the term is often used in human factors area to express a phenomenon that the human does not monitor the automation. Moray and Inagaki [17] have pointed out that the usage is misleading, because 'not monitoring the automation' does not necessarily mean that the human is complacent. An obvious counterexample is that the human is busily occupied with extremely urgent tasks. Therefore this paper tries to avoid using the term complacency.

This paper proposes a theoretical framework to describe, analyze, and evaluate

the driver's overtrust in and overreliance on ADAS. Although the two notions, overtrust and overreliance, are often used as if they are synonyms, this paper differentiate the notions rigorously. To this end, two aspects, (1) situation diagnostic aspect and (2) action selection aspect, are introduced. The first aspect is to describe overtrust, and has three axes: (1-1) dimension of trust, (1-2) target object, and (1-3) chances of observation. The second aspect, (2), is to describe overreliance on the ADAS, and distinguishes the following two types of decisions: (2-1) commission like action selection decision and (2-2) omission like action selection decision.

2 OVERTRUST

Overtrust can be defined as a psychological state in which the human trust is inappropriately high. Overtrust is an incorrect situation diagnostic decision claiming that the object is trustworthy when it actually is not. This paper introduces three axes for describing the types of overtrust in a precise manner.

2.1 Dimension of trust

The first axis (1-1) gives the dimension of trust. Lee and Moray [18] have distinguished four dimensions for trust: (a) foundation, representing the fundamental assumption of natural and social order, (b) performance, resting on the expectation of consistent, stable, and desirable performance or behavior, (c) process, depending on an understanding of the underlying qualities or characteristics that govern behavior, and (d) purpose, resting on the underlying motives or intents. Trust in an object is appropriate when all the dimensions are evaluated correctly. When there is a dimension that is evaluated inappropriately high, perceived trust is seen as overtrust. Therefore some types can be distinguished for overtrust depending on which dimension of trust is violated.

Example 1: Suppose the driver thought that, "The ADAS has been successful in coping with the situations so far. The system will continue to be successful hereafter, too." This is a type of overtrust, violating the second dimension of trust.

Example 2: Imagine a case in which the driver thought that, "I do not know how the function is implemented in the ADAS. I am not informed how the task is carried out, either. However, it would be quite alright even if I do not know the details." This is a type of overtrust, violating the third dimension of trust.

Example 3: Assume that the driver said that, "I do not understand why the system is doing such a thing. However, the system should be doing what it thinks it necessary and appropriate. The system will not harm us." This type of overtrust does not satisfy the forth dimension of trust.

2.2 Target object to which overtrust is addressed

The second axis (1-2) describes a target object to which the driver places inappropriately high trust. This paper distinguishes five types of target objects, computer (C), software (S), hardware (H), environment (E), and liveware (L) according to the C-SHEL model [19] describing human interactions with other humans, technology, and the environment; see, Figure 1.
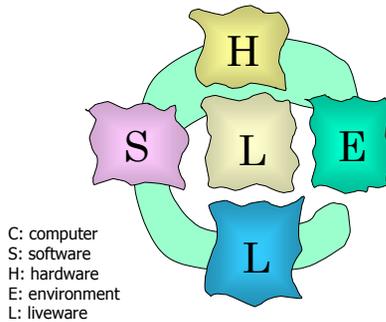
C: computer
S: software
H: hardware
E: environment
L: liveware

Fig. 1. C-SHEL model

Example 4 (overtrust in computer): The adaptive cruise control system (ACC) performs the longitudinal control on behalf of a driver. Suppose the driver thought that, "A car just ahead of me on the next lane may be cutting in. The ACC must have already noticed the car and will adjust the control when appropriate." This is overtrust in the ACC (computer) when the car on the next lane is outside the range of the ACC and the driver does not notice that.

Example 5 (overtrust in software): Imagine a case in which a driver thought that, "Today is the first day for me to use a brand new system. Oh, I forgot to read the manual. There should be no problem even if I pressed the buttons in a wrong sequence. Fool-proof or tamper-proof functions must be implemented in the software."

Example 6 (overtrust in hardware): Assume that a driver thought that, "Strictly speaking, this is the time for me to bring the car to a periodic inspection. However, I am quite busy right now and I have never experienced hardware troubles in the car. Why do I have to bring my car periodically for an inspection? My car will not fail."

Example 7 (overtrust in environment): Suppose a man is driving his car thinking that, "This road is simply straight. Moreover, there is usually little traffic. It is very relaxing to drive on this simple and somewhat boring road." In reality, environment may alter with time.

Example 8 (overtrust in liveware):   Suppose a driver is approaching to an intersection with a blind corner and that an ADAS sets off an alert telling the driver that, "A car is approaching to the intersection from the right on the crossing road." The driver cannot see the car himself, because the car is just behind the blind corner. The ADAS generated the alert based on the information obtained via vehicle-to-vehicle or vehicle-to-infrastructure communication technology. Suppose the driver thinks that, "I do not see any car. If there is a car, the car will surely yield the right of way, because it is I that is on the priority road," which is overtrust in the driver (liveware) of the other car at the blind corner.

Example 9 (overtrust in liveware): Imagine a car equipped with an electronic stability control system (ESC) that improves stability by applying the brakes to

individual wheels when skids or loss of steering control was detected. Suppose the ESC worked at a sharp curve on a slippery road. If the human interface was not properly designed to let the driver know the ESC was activated, the driver might feel inappropriate confidence on his driving skill, failing to recognize that it was the ESC that assured the stability of the car at the curve. This is a case of overtrust in the driver himself/herself.

2.3   Chances of observation

The third axis (1-3) distinguishes two classes for ADAS: (a) ADAS for use in normal driving and (b) ADAS for use in emergency. A most prominent characteristic that distinguishes the two classes is the chances to observe ADAS functioning.

Example 10: ADAS for use in normal driving (e.g., ACC) usually aims to reduce the driver workload, and works continuously for certain period of time. Since such an ADAS is used daily, the driver observe the system's 'intelligent' behaviors repeatedly, which gives the driver a number of opportunities for constructing a mental model of the ADAS.

Example 11: ADAS for use in emergency (e.g., PCS described in section 1) usually aims to prevent a catastrophic event from occurring and thus to attain the driver safety. Since such an ADAS is activated only in cases of emergency, it would be very rare for an ordinary driver to see the ADAS works. That suggests that the driver may not be able to accumulate chances sufficient enough for constructing a concrete mental model of the ADAS.

3   OVERRELIANCE

Overreliance on an ADAS is a psychological state in which the human reliance on the ADAS is inappropriately high. Overreliance is an incorrect action selection decision based on an incorrect situation diagnostic decision on the ADAS (i.e., the overtrust in it). For the action selection, (2), this paper distinguishes the following two types of decisions: (2-1) commission like action selection decision and (2-2) omission like action selection decision. The former is a selection and implementation of an action that is not suitable to a given situation. The latter is a failure to select or implement an action that is needed in a given situation.

Example 12 (commission like action): Suppose a man is driving a car equipped with an ESC at high speeds, which is overreliance on the ESC, if it was a clear but extremely cold winter morning and it had rained before dawn. It would be inappropriate to drive a car at high speeds in such an adverse weather condition although the car is equipped with the ESC.

Example 13 (omission like action): Suppose a man is driving a car by using an ACC and a lane keeping assistance system (LKA). LKA is an automatic system that recognizes the lane and provides the driver with assisting steering torque to keep the car around the center of the lane. Suppose the driver decided to let the LKA take care of the lateral control completely for a while so that he could consult the navigation system to know how to access his destination. If the LKA was of the type that ceases to control the steering when it determines, through monitoring the driver behavior, that the driver has not been active in steering,

the driver's decision to trade the full authority to the LKA is overreliance on the LKA. A case may happen that nobody controls the car, if the human interface did not tell the driver clearly that the LKA returned the authority and responsibility of steering back to the driver based on its decision that the driver had been inactive in steering for certain period of time.

4   POSSIBILITIES OF OVERTRUST AND OVERRELIANCE

Let us discuss overtrust in and overreliance on ADAS by integrating viewpoints given in sections 2 and 3.

4.1   Communication-based information provision

Suppose an ADAS has a communication-based function to set off an alert on a car that the driver may not be able to see. There are some objects in which the driver may place overtrust. Example 8 has described one of such cases, where the driver of some other car (a liveware in the target-object-axis in section 2.2) needs to be taken into account from a viewpoint of performance in the dimension-of-trust-axis in section 2.1.

Consider a case in which a driver is approaching to an intersection that has blind corners but has no traffic lights. The communication-based infrastructure was installed a year ago. The infrastructure can detect cars travelling on the roads crossing each other, and it sends a signal to an onboard ADAS of a car so that the ADAS can set off an alert to let the driver know an approach or existence of some car(s) on a crossing road. Suppose the driver drives the road daily (i.e., chance-of-observation axis) and has been satisfied with the performance (i.e., dimension-of-trust-axis) of the communication-based alert. The driver now thinks that, "I am sure that no car is coming toward me when no alert is given. Why not cross the intersection without deceleration?" In this case, the driver is overlooking the possibility of hardware failure of the infrastructure (i.e., target-object-axis). His situation-diagnostic decision that, "No car must be approaching toward me because no alert is there" is inappropriate (i.e., overtrust). When the communication-based infrastructure was out of service, no alert can be given to the driver. Thus the action selection decision to "cross the intersection without deceleration" is overreliance on the function of the communication- based alert, when the driver abandons the responsibility to be vigilant.

4.2   Adaptive cruise control system

Itoh [20] has found an example of overtrust in and its resulting overreliance on the ACC through his experiment with a driving simulator. Participants were requested to drive a car by using an ACC that can control the host vehicle to a complete stop when the lead vehicle decelerates and stops. However, the ACC does not recognize stationary body (such as, cars standing still). Participants experienced 69 drives with ACC during the period of four days. At the final trial on the fourth day, participants were given a case in which, after 20 minutes of following the lead vehicle at 100km/h, the lead vehicle made a lane change and the host vehicle happened to approach to the tail of a traffic jam, where all the vehicles in the jam stood completely still. Participants needed to apply the brake by themselves. One collision and some near-collisions into the car at the tail of

the jam were observed in the experiment. None of the participants who caused the collision or near-collision were drowsy or distracted. Data analyses and investigation of those cases suggested that the participants developed trust in the ACC while experiencing repeatedly the ACC's successful lead vehicle followings to complete stops (i.e., chance-of-observation-axis), and that some participants had inappropriate expectations (i.e., dimension-of-trust-axes) that the "ACC would control the host vehicle nicely to a vehicle ahead," even for an already standing still vehicle. The participants' failure in applying the brake (i.e., omission like action) is due to overreliance on the ACC, induced by the overtrust in it.

4.3   Pre-crash safety system

Situation would be quite different in cases of ADAS for emergency, such as PCS. The PCS usually aims not to prevent a catastrophe from occurring but to mitigate collision damages. Since the system is activated only in cases of emergency, it would be very rare for an ordinary driver to see how the system works (i.e., chance-of-observation-axis). It is thus highly possible that the driver will not be able to construct a precise mental model of the PCS. This suggests that it may be hard for the driver to engender a sense of trust in the system (i.e., dimension-of-trust-axis). What happens then? No possibility for the driver to place overtrust in the PCS? The answer may be negative. It is known that people may place inappropriate trust (i.e., overtrust) without having any concrete experience or evidence proving that the object is trustworthy; see, e.g., [18].

Suppose the driver places overtrust in the system. Does that mean that the driver relies on the system (i.e., overreliance)? The answer may be negative again. In case of an ADAS designed for use in normal driving situations, even if the system's behavior was not what the driver expected, there would be enough time for the driver to override the system to cope with the circumstances himself or herself. However, in case of an ADAS for emergency use, even if the driver noticed that the system's behavior was not what he or she expected, no time may be left for him or her to correct it. Even so, does the driver rely on the ADAS (i.e., overreliance) and allocate his or her resources to something else at the risk of his or her life?

5   DISCUSSIONS

This paper has proposed a theoretical framework to discuss the driver's overtrust in and overreliance on ADAS in a precise manner. Overtrust and overreliance are distinguished rigorously and their characteristics are illustrated by introducing some viewpoints (or, aspects and axes). It has been shown that our theoretical frame enables precise description and classification as well as rigorous analysis and evaluation of the driver's overtrust in and overreliance on ADAS. Since the framework distinguishes the target object of the driver's overtrust, it can be used to derive a countermeasure for reducing the possibilities of the driver's overtrust. In other words, a systematic investigation can be made possible to determine whether a type of overtrust in question may be alleviated by improving human-machine interface, or by preparing a better operation manual, or by providing the drivers with opportunities to acquire

knowledge and/or to improve skills, or by some other means.

It would be apparent that alleviation or prevention of overtrust in or overreliance on the ADAS and its effects on degradation of safety of the car-driver system is closely linked to the issue of authority and responsibility. It is sometimes useful to provide the driver with multi-layered assist functions [1]. In the first layer, a driver's situation recognition and understanding are enhanced for proper situation diagnostic decisions and associated action selection decisions. In the second layer, the ADAS monitors the driver's behaviours and traffic conditions to evaluate whether his or her intent and behaviours match the traffic conditions. When the ADAS detects a deviation from normality (for instance, by detecting behaviours or postures that suggest the driver's overtrust or its resulting overreliance), it gives the driver an alert to make him or her return to normality. In the third layer, the ADAS provides the driver with automatic safety control functions, if the deviation from normality still continues to be observed or if little time is left for the driver to cope with the traffic conditions. The situation-adaptive ADAS adjusts its assist functions dynamically so that they may fit to the human's intent, psychological/physiological conditions, and the traffic conditions. The adjustment of assist functions is made in a machine-initiated manner [21-23] by inferring intent and conditions of the human through monitoring his or her behaviours.

The driver's control action may be classified into three categories: (1) An action that needs to be done in a given situation, (2) an action that is allowable in the situation and thus it may either be done or undone, and (3) an action that is inappropriate and thus must not be done in the situation. Assuming sensing technology for the computer (ADAS), two states may be distinguished for each control action: (a) "Detected," in which the computer judges that the driver is performing the control action, and (b) "undetected," in which the control action is not detected by the computer (Figure 2). Region A represents the cases of the driver's omission like action selection, while Region B depicts the cases of the driver's commission like action selection and implementation. These mismatches between the driver's action selection decision and the given situation can occur when the driver may place overreliance on the ADAS, as has been discussed already. Then the question becomes, "What is a sensible and effective countermeasure for the ADAS in such circumstances? Is it enough for the ADAS to set off an alert to let the driver resolve the mismatch himself or herself? Or, is it better for the ADAS to initiate an automatic control action to cope with the situation?" Inagaki and his colleagues have shown that the authority may be given to the ADAS so that it can take an automatic safety control action that the driver failed to perform or can take a protective action (soft protection or hard protection) that tries to prevent the driver's inappropriate action causing an accident or an incident [24-26].
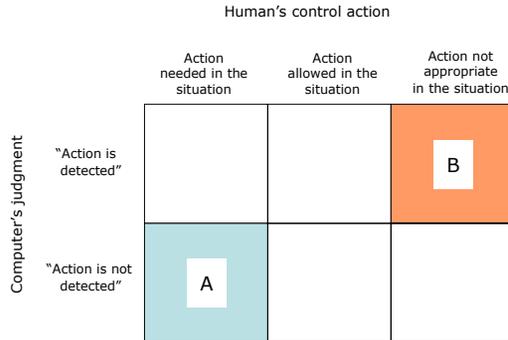
Human's control action

	Action needed in the situation	Action allowed in the situation	Action not appropriate in the situation
"Action is detected"			B
"Action is not detected"	A		

Computer's judgment

Fig. 2. Control action in a given situation

6 REFERENCES

[1] Inagaki, T. (2008). Smart collaborations between humans and machines based on mutual understanding. Annual Reviews in Control, vol. 32, pp. 253-261.

[2] Hollnagel, E. & Bye, A. (2000). Principles for modeling function allocation. Int. J. Human-Computer Studies, 52, 253-265.

[3] Klein, G. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. Klein, et al (Eds.), Decision making in action (pp. 138-147). Ablex.

[4] Billings, C.E. (1997). Aviation automation – The search for a human-centered approach. LEA.

[5] Inagaki, T. (2006). Design of human-machine interactions in light of domain-dependence of human-centered automation. Cognition, Technology & Work, 8(3), pp. 161-167.

[6] Wickens, C.D. (1994). Designing for situation awareness and trust in automation. Proceedings of IFAC Integrated Systems Engineering, 77-82.

[7] Endsley, M.R. & Kiris, E.O. (1995). The out-of-the-loop performance problem and the level of control in automation. Human Factors, 37(2), 3181-3194.

[8] Sarter, N.B. & Woods, D.D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. Human Factors, 37(1), 5-19.

[9] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230-253.

[10] Sarter, N.B., Woods, D.D., & Billings, C.E. (1997). Automation surprises. In G. Salvendy (Ed.). Handbook of human factors and ergonomics (2nd ed., pp. 1926-1943). Wiley.

[11] Inagaki, T., & Stahre, J. (2004). Human supervision and control in engineering and music: Similarities, dissimilarities, and their implications.

Proceedings of the IEEE, 92(4), 589-600.

[12] Parasuraman, R., Molloy, R. & Singh, I.L. (1993). Performance consequences of automation-induced 'complacency.' Int. J. of Aviation Psychology, 3(1), 1-23.

[13] Mosier, K., Skitka, L.J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. International J. Aviation Psychology, 8, 47-63.

[14] Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. Human Factors, 43, 563-572.

[15] Sheridan, T.B., & Parasuraman, R. (2005). Human-automation interaction. In. R.S. Nickerson (Ed.). Reviews of Human Factors and Ergonomics, Volume 1 (pp. 89-129). HFES.

[16] Complacency (2010). In Merriam-Webster Online Dictionary.

[17] Moray, N., & Inagaki, T. (2001). Attention and complacency. Theoretical Issues in Ergonomics Science, 1(4), 354-365.

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

[19] Inagaki, T. (2003). New challenges on vehicle automation: Human trust in and reliance on adaptive cruise control systems. Proc. IEA 2003 (CD-ROM), 4 pages.

[20] Itoh, M. (2009). Contributing factors to driver's over-trust in a driving support system for workload reduction. Trans. SICE, 45(11), 555-561 (in Japanese).

[21] Inagaki, T. (2003). Adaptive automation: Sharing and trading of control. In E. Hollnagel (Ed.) Handbook of cognitive task design (pp. 147-169). LEA.

[22] Scerbo, M. W. (1996). Theoretical perspectives on adaptive automation. In R. Parasuraman & M. Mouloua (Eds.). Automation and human performance (pp.37-63). LEA.

[23] Inagaki, T., & Sheridan, T.B. (2008). Authority and responsibility in human-machine systems: Is machine-initiated trading of authority permissible in the human-centered automation framework? Proc. Applied Human Factors and Ergonomics 2008 (CD-ROM) 10 pages.

[24] Inagaki, T., Itoh, M., & Nagai, Y. (2006). Efficacy and acceptance of driver support under possible mismatches between driver's intent and traffic conditions. Proc. HFES 50th Annual Meeting, 280-283.

[25] Inagaki, T., Itoh, M., & Nagai, Y. (2007). Driver support functions under resource-limited situations. Proc. HFES 51st Annual Meeting. 176-180.

[26] Inagaki, T., Itoh, M., & Nagai, Y. (2007). Support by warning or by action: Which is appropriate under mismatches between driver intent and traffic conditions?. IEICE Trans. Fundamentals. E90-A(11), 264-272.