Appeared in *Cognition Technology & Work*, 14(1), 29-37, 2012.

Authority and Responsibility in Human-Machine Systems:

Probability theoretic validation of machine-initiated trading of authority

Toshiyuki INAGAKI

Department of Risk Engineering
University of Tsukuba
Tsukuba 305-8573 JAPAN
E-mail: inagaki.toshiyuki.gb@u.tsukuba.ac.jp

Thomas B. SHERIDAN

Departments of Mechanical Engineering and Aeronautics / Astronautics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139-4307 USA
E-mail: Sheridan@mit.edu

Correspondence to:

T. Inagaki
Department of Risk Engineering, University of Tsukuba, Tsukuba 305-8573, JAPAN
Email: inagaki.toshiyuki.gb@u.tsukuba.ac.jp
Phone & Fax: +81-29-853-5537

Abstract: Human-centered automation is an approach to realize a work environment in which humans and machines cooperate. It is usually claimed in the framework that “the human must have final authority over the automation.” However, correctness of the statement is context-dependent: we note that humans have limited capabilities and authority is interconnected with responsibility. This paper illustrates the need for a machine-initiated trading of authority from humans to automation in the vehicle driving context, and clarifies issues to be solved for implementing useful automation invocation based on the machine’s interpretation of the situation and the human’s behavior.

Keywords: Authority and responsibility; adaptive automation; behavior monitoring; human-centered automation; systems safety; trading of authority;

1 Introduction

Many current systems are semi-autonomous. Humans are usually assumed to be responsible for the safety of the human-machine systems, and thus are considered to be in command in those systems (see, e.g., Woods 1989; Billings 1997). If a human’s decision and its associated directive to the machine are correct, the obtained result will match the human goal and the situation. In reality, however, a human can fail to give a proper directive to the machine. A most obvious case may be where the human’s situation awareness (SA) is inappropriate or incomplete. When a human’s SA is inappropriate or incomplete, the human’s decision and action are likely to be incorrect. Three levels are often distinguished for situation awareness: Level 1 SA is defined as “perceiving critical factors in the environment,” Level 2 SA as “understanding what those factors mean, particularly when integrated together in relation to the person’s goal,” and Level 3 SA as “understanding of what will happen with the system in the near future” (Endsley 1996). Some causes for failure in attaining Level 1 SA are inattention, internal/external distractions, and improper observation. When a human lacks the proper Level 1 SA, he/she may fail to notice that some control action is needed in the situation. Level 2 SA may be lost for reasons such as misjudgment (based on incorrect or incomplete knowledge) or a false assumption of the given situation. When this is the case, the human may select an inappropriate control action that does not fit the given situation. There are many factors that can prevent the human from attaining Level 3 SA. Poor knowledge and imprecise information about the system, and incorrect risk perception of the environment are some such factors. When a human’s Level 3 SA is incorrect, the human may fail to take the necessary control action at the right moment.

It is noted that correct SA by a human does not assure that an undesirable result can be avoided. Even when the human's SA is completely correct, if allowable time is limited, he/she may fail to take a necessary countermeasure or to give a proper directive to the machine. Suppose an accident occurred in a system. People sometimes blame the human operator of the system. If the judgment is that the human operator failed to interpret the situation, to predict what would happen in the near future, or to take a necessary control action for avoiding an unwanted result, the operator may be accused legally. If there was an extremely short time allowance that hindered a human's decision and control, it may not be wise to blame the human by assuming that that human is responsible for the system's safety just because every authority was given to him/her.

The operating environment can change as time passes, and the human's performance may degrade due to psychological/physiological reasons. Moreover, humans have limited capabilities. Today's machines can sense and analyze a situation, decide what must be done, and implement control actions. Should such an intelligent machine do nothing if it is not given a directive by a human, even when it has detected that the human is late in taking a control action that is needed in a given situation? Should such an intelligent machine sit back when it detects a human's apparently inappropriate control action, by assuming that the human must have some good reason for doing so? Allowing a machine to take a corrective control action when it believes that the human is late in taking a necessary measure or behaving inappropriately implies that the authority is traded from the human to the machine temporarily. Strategies for trading of authority are classified into two disjoint groups; viz., *human-initiated* strategies and *machine-initiated* strategies (Scerbo 1996). In the former, the human is in command. However, the human is not in command in the latter, at least for a while, because the need of and the timing of automation invocation is decided based on the judgment of the machine and is implemented without any human intervention. In other words, machine-initiated strategies do not comply with the principles of human-centered automation (Billings 1997; ICAO 1998). Whether machine-initiated strategies are permissible or not is one of the crucial issues in adaptive automation (see, e.g., Rouse 1988; Parasuraman, Bhari, Deaton, Morrison, & Barnes 1992; Scallen & Hancock 2001; Scerbo 1996; Inagaki 2003).

Based on probability theoretic analyses, this paper argues that a machine-initiated trading of authority may be indispensable even in the framework of human-centered automation when safety of human-machine systems is a major factor.

2 Mismatches between actions and given situations

A human's control action or the human's directive to the machine may be classified into three categories: (1) A control action that needs to be done in a given situation, (2) a control action that is allowable in the situation and thus may either be done or not done, and (3) a control action that is inappropriate and thus must not be done in the situation. Assuming some sensing technology (or machine intelligence, provided by a computer), two states may be distinguished for each control action: (a) "Detected," in which the computer judges that the human is performing the control action, and (b) "Undetected," in which the control action is not detected by the computer. Figure 1 depicts all possible combinations of a control action and its state. Among them, case α shows a circumstance in which the computer judges that the human operator is (too) late in performing or ordering a control action that must be done in the given situation. A typical example of case α in the automobile domain is that, in spite of a rapid deceleration of a lead vehicle, a car driver does not apply the brakes due to some distraction. Case β indicates a circumstance in which the computer determines that the human operator misunderstands a given situation and the control action that he/she takes or requests does not fit the situation. A typical example of case β is that a driver is about to steer the wheel to enter into an adjacent lane without noticing that a faster vehicle is approaching from behind on the lane.

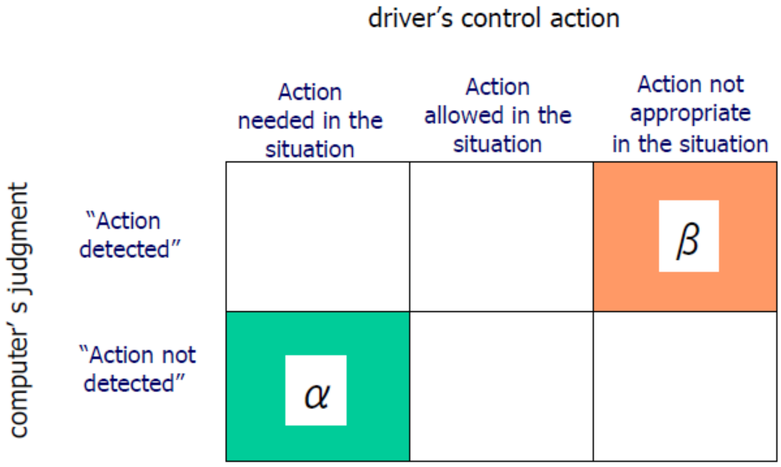
driver's control action
Action needed in the situation
Action allowed in the situation
Action not appropriate in the situation
computer' s judgment
"Action detected"
"Action not detected"
β
α

Figure 1: Control action in a given situation

A question that must be asked for case α is whether the computer may be allowed to initiate without human intervention the control action (such as to apply the brakes) that the human should have done, or whether the computer is allowed only to set off a warning to urge the human to perform manually the control action that the situation requires. A question asked for case β is whether the computer may be allowed to prohibit the control action (such as steering the wheel to make a lane change) that the human is trying to do, or whether the computer is allowed only to set off a warning to tell the human that his/her action should be stopped at once.

Suppose the computer always knows what control action is appropriate beyond just detecting (or not) whether a control action is taken. Then it would be almost obvious that the computer may be allowed to initiate the control action that the human failed to perform in case α, and that the computer may be allowed to prohibit the human's control action that does not fit a given situation in case β, considering the following facts: (1) humans do not always respect or respond to warnings and (2) humans need some amount of time to interpret the warnings and time delay is inevitable until effective actions are taken. However, it is too optimistic to assume that the computer never makes an error in judging whether the human's response to the situation is inappropriate. By taking examples from the automobile domain, sections 3 and 4 analyze the efficacy of the computer's support for cases α and β, respectively, under a realistic setting that the computer's judgment may be wrong.

3 Support by warning or by action: Case α, braking

Let us define case α as follows: The computer detects a rapid deceleration of a lead vehicle. In addition, it notices that the driver of the own vehicle has not applied the brakes yet. The computer thus determines that some support needs to be given to the driver.

3.1 Model

The computer has a *situation monitoring* function to monitor the traffic situation as well as a *behavior checking* function to detect the driver's context-specific control action to cope with the situation. Let the true state of the traffic situation be represented by U for *unsafe* (e.g., a rapid deceleration is made by the lead vehicle) and S for *safe* (e.g., no such a rapid deceleration is there). The computer judges whether the state of the traffic situation is unsafe (denoted "U") or safe (denoted “S”). The computer’s situation monitoring capability is imperfect and is represented by the conditional probabilities P(“U”|U), P(“S”|U), P(“U”|S), and P(“S”|S). Behavior checking is activated only when the computer determines that the traffic situation is unsafe (“U”). Let the true state of the driver be represented by A for a control *action* (e.g., braking) and NA for *no action*. The computer determines, through behavior checking, whether the driver is taking the control action (denoted “A”) or not (denoted “NA”). The computer’s behavior monitoring capability is also imperfect and is expressed by P(“A”|A), P(“NA”|A), P(“A”|NA), and P(“NA”|NA).

The computer gives support to the driver according to the following rules:

(i) If the computer determines that the traffic situation is unsafe (“U”) and that the driver is taking no control action (“NA”) to cope with the situation, then it gives a support to the driver.

(ii) If the computer finds the situation unsafe (“U”) and that the driver applying a proper control action (“A”), then it does not intervene into what the driver is doing.

(iii) If the situation is regarded as being safe (“S”), the computer does not perform the driver behavior checking (Fig.2).

Two kinds of support are distinguished for case α: One is the *support by warning* in which a warning is set off to urge the driver to apply the brakes, and the other the *support by action* in which emergency brakes are applied automatically. It is assumed that the computer can issue a warning or apply emergency brakes properly when it intended to do so. In other words, issuing a warning and applying emergency brakes are likely to be performed highly reliably by the automation.

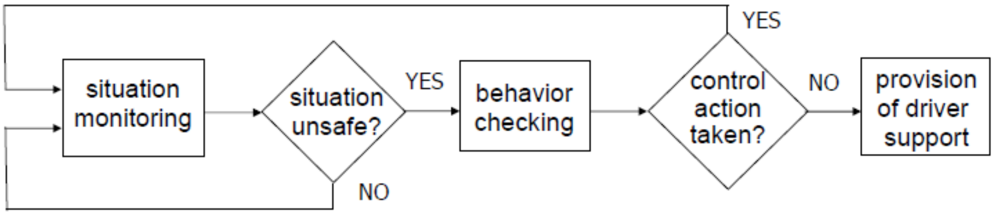

Figure 2: Situation monitoring followed by behavior checking

Let the *state of the world* for case α be defined as (state of the traffic environment, state of the driver). Characteristics of the support are evaluated from three viewpoints: (a) When the true state of the world is (U, NA) in which the driver fails to take any action to cope with the unsafe situation, to what extent can the support avoid an accident properly? (b) When (U, A) in which the driver is already responding to the unsafe situation, should the computer interfere with the driver by providing unnecessary support? (c) When (S, NA) in which the situation is safe, would the computer cause inconvenience by providing wrong and inappropriate support?

3.2 Probabilistic evaluation of support by warning

(a) Accident prevention when (U, NA)

A warning is issued when (U, NA) if the computer finds the situation unsafe (“U”) and determines that the driver is taking no action (“NA”) to cope with the situation. Let P(“U”, “NA”|U, NA) denote the conditional probability that a correct warning is set off when (U, NA). Under the assumption that the computer’s functions for situation monitoring and behavior checking can fail statistically independently, we have: P(“U”, “NA”|U, NA) = P(“U”|U) P(“NA”|NA). Whether an accident can be avoided or not depends on whether the driver respects a warning and initiates an action to cope with the situation. Let P(IA|warning) denote the conditional probability that the driver initiates such an action (IA) upon receiving a warning. The conditional probability of accident prevention when (U, NA) is given by:

$$P_w(\text{accident prevention}|\text{U, NA}) = P(\text{“U”}|\text{U})\ P(\text{ “NA”}|\text{NA})\ P(\text{IA}|\text{warning}) \qquad (1)$$

where $P_w(\cdot|\cdot)$ denotes the conditional probability under support by a warning (whether appropriate or not).

(b) Unnecessary support when (U, A)

Warning the driver to request a control action that he/she is already performing to cope with the unsafe situation may cause driver confusion or annoyance. The computer issues such an unnecessary warning when (U, A) if its correct understanding of the situation (“U”|U) is followed by a miss (“NA”|A) of the driver’s proper control action to cope with the unsafe situation. The conditional probability of an unnecessary warning when (U, A) is given by:

$$P_w(\text{unnecessary warning}|\text{U, A}) = P(\text{“U”}|\text{U})\ P(\text{ “NA”}|\text{A}) \qquad (2)$$

(c) Wrong and inappropriate support when (S, NA)

The state (S, NA) does not require any warning. A wrong and improper warning is set off when (S, NA) if the computer's incorrect situation interpretation ("U"|S) is followed by correct driver behavior checking ("NA"|NA). Then we have:

$$P_w(\text{inappropriate warning}|S, NA) = P(\text{"U"}|S)\, P(\text{"NA"}|NA) \tag{3}$$

3.3 Probabilistic evaluation of the support by action

(a) Accident prevention when (U, NA)

The conditional probability that an accident can be prevented from occurring when (U, NA) is given by:

$$P_a(\text{accident prevention}|U, NA) = P(\text{"U"}|U)\, P(\text{ "NA"}|NA) \tag{4}$$

where $P_a(\cdot|\cdot)$ denotes the conditional probability under the support by an action (whether appropriate or not). It is seen, from (1) and (4), that support by action is always more effective than the support by warning in avoiding an accident, since the driver may not heed the warning.

(b) Unnecessary support when (U, A)

The computer applies automatic emergency brakes when (U, A) if its correct situation interpretation ("U"|U) is followed by a missed detection ("NA"|A) of the driver's braking action. The conditional probability of unnecessary automatic braking when (U, A) is given by:

$$P_a(\text{unnecessary automatic braking}|\ U, A) = P(\text{"U"}|U)\, P(\text{ "NA"}|A) \tag{5}$$

The automatic brakes are 'unnecessary' in the sense that they are redundant to the driver's manual braking.

(c) Wrong and inappropriate support when (S, NA)

Automatic emergency brakes are applied inappropriately when (S, NA) if the computer's incorrect situation interpretation ("U"|S) is followed by correct driver behavior checking ("NA"|NA). The conditional probability of wrong and improper automatic braking when (S, NA) is given by:

$$P_a(\text{inappropriate automatic braking}|S, NA) = P(\text{"U"}|S)\, P(\text{ "NA"}|NA) \tag{6}$$

3.4 Observations

(I) Suppose the computer makes no errors in situation monitoring (P(“U”|U) = P(“S”|S) = 1) and in driver behavior checking (P(“A”|A) = P(“NA”|NA) = 1). Then the *support by action* (SBA) provides better performance than the *support by warning* (SBW). This is because:

(i) neither SBA nor SBW offers unnecessary support when the driver is already responding to the situation, equations (2) and (5);

(ii) neither SBA nor SBW activates inappropriate support under the safe situation, equations (3) and (6); and

(iii) the SBA never fails to avoid an accident, while the SBW can avoid an accident only when the driver responds to a warning in a timely and proper manner, equations (1) and (4).

(II) Suppose the computer may make errors in the situation monitoring (P(“U”|U) $\leq$ 1, P(“S”|S) $\leq$ 1) and/or in the driver behavior checking (P(“A”|A) $\leq$ 1, P(“NA”|NA) $\leq$ 1). The SBA’s primacy over the SBW is still maintained in accident avoidance. However, even the SBA cannot always avoid an accident because P_a(accident prevention|U, NA) $\leq$ 1.

(III) The imperfection of the computer’s situation monitoring and/or driver behavior checking functions brings problems to both SBA and SBW. The first problem is an unnecessary support given when (U, A), which is due to missed detection P(“NA”|A) of the driver’s control action. Although the probability of such an unnecessary support is the same between SBA and SBW, the outcome differs slightly between these two. The SBW’s warning issued when (U, A) requesting the driver to initiate an action that he/she is already taking may be confusing or annoying. The SBA’s automatic emergency braking applied when (U, A) may make the driver feel that his/her braking might be a bit stronger than expected. In reality, the missed detection P(“NA”|A) may rarely happen for case α because it is not difficult for the computer to judge correctly whether the driver is applying the brakes or not, a capability already inherent in electronic control of today’s automobile.

(IV) The second problem is a wrong and inappropriate support given when (S, NA), which is due to the computer’s incorrect interpretation of the situation, P(“U”|S). The resulting inconvenience is more serious in the SBA than in the SBW. The SBA’s automatic emergency brakes in the safe situation may bring a chance to be rear-ended. A wrong and inappropriate warning when (S, NA) is also a crucial issue for the SBW, because the driver’s distrust of the system needs to be avoided. How can we make P(“U”|S) smaller? One simple way might be to give the computer time to collect more information and judge the state of the situation at the latest time possible. However, that may bring disadvantages to the SBW. For a warning given later it becomes harder for the driver to react to the warning and implement a required control action within the shorter available time.

4 Support by warning or by action: Case β, lane changing

Let us define case β as follows: While monitoring the driver’s behavior, the computer detects the driver’s lane change intent, say, through the increase in the frequency of visual glances in the side mirror (see, e.g., Zhou, Itoh, & Inagaki 2009). However, it notices that a faster vehicle is approaching from behind on the same lane. The computer thus determines that some support must be given to the driver.

4.1 Model

By use of its *behavior monitoring* function, the computer monitors the driver’s behavior to detect any control action that can change the vehicle from a stable and safe state into a new and possibly unsafe one. Let the true state of the driver be represented by A for a control *action* (e.g., making a lane change), and NA for *no action*. The computer determines whether the driver is about to initiate such an action (denoted “A”) or not (denoted “NA”). If a control action (or the driver's intention to initiate the control action) is detected (“A”), the computer activates the *situation checking* function, to validate the adequacy of the control action in the traffic environment. Let the true state of the traffic environment be represented by S for *safe* with respect to the control action (e.g., vehicles are far behind on the adjacent lane) and U for *unsafe* (e.g., a faster vehicle is approaching from behind on the adjacent lane). The computer judges whether the traffic environment is safe (denoted “S”) or unsafe (denoted “U”) for the detected control action. It is assumed that the computer may be wrong in behavior monitoring and/or situation checking.

The computer gives support to the driver according to the following rules:

(i) If the computer detects that the driver is about to initiate a control action (“A”) and determines that the action does not fit the situation (“U”), then it gives support to the driver.

(ii) If the computer detects the driver's control action ("A"), if it judges that the action fits the situation ("S"), then it does not intervene into the driver's action.

(iii) If the computer finds no action by the driver ("NA"), then it continues behavior monitoring (Fig. 3). Note here that the behavior monitoring is followed by the situation checking in Fig. 3, while the situation monitoring was followed by the behavior checking in Fig. 2. In other words, the order to use the functions for understanding the driver's behavior and the traffic situation is quite opposite between cases α and β.

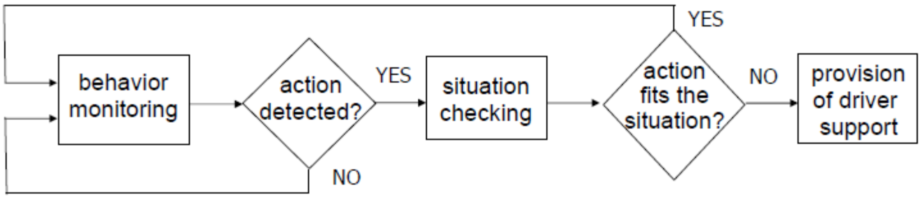

Figure 3: Behavior monitoring followed by situation checking

The computer's support to the driver is (again) categorized into two classes: One is the *support by warning* that sets off a warning telling the driver that the action he/she is about to take does not fit the situation. The other is the *support by action* that performs protective control to prohibit the driver's action from being implemented. The support by action is further divided into two subclasses: *Hard protection* where the driver's control input is neglected for a while and he/she is not allowed to override the computer's protective control, and *soft protection* where the steering wheel becomes slightly stiffer than usual but the driver may override the computer's protective control by adding more force into the steering wheel when he/she thinks it necessary. It is assumed that the computer can issue a warning or apply either type of protection properly and reliably when it intended to do so.

Let the state of the world for case β be defined as (state of the driver, state of the traffic environment). Note here (again) that the order of the constituents of the state of the world is reversed here, compared to case α, in order to reflect the fact that the state of the traffic situation is checked when the driver's control action is detected. Capabilities of the support are evaluated from three viewpoints: (a) When the true state of the world is (A, U) in which the driver’s action does not fit the situation, to what extent can the support properly avoid an accident? (b) When (A, S) in which the driver’s action fits the situation, to what extent can the computer interfere with the driver by providing wrong and inappropriate support? (c) When (NA, S) or (NA, U) in which the driver is not initiating an action, to what extent would the computer provide unnecessary support?

4.2 Probabilistic evaluation of the support by warning

(a) Accident prevention when (A, U)

A warning is set off when (A, U) if the computer detects the driver’s lane change action (“A”) and finds that the action does not fit the situation (“U”). Under the assumption that behavior monitoring and situation checking fail statistically independently, P(“A”|A) P(“U”|U) gives the conditional probability that a correct warning is set off when (A, U). Whether an accident can be avoided or not depends on whether the driver respects the warning and abandons his/her intention for a lane change. Let P(BO|warning) denote the conditional probability that the driver breaks off (BO) his/her intended action upon receiving a warning. The conditional probability that an accident (colliding into a faster vehicle from behind on the adjacent lane) is avoided when (A, U) is given by:

$$P_w(\text{accident prevention}|A, U) = P(\text{“A”}|A)\ P(\text{ “U”}|U)\ P(BO|\text{warning}) \qquad (7)$$

(b) Wrong and inappropriate support when (A, S)

Suppose a lane change that the driver intends is safe and proper. In that situation warning the driver not to make a lane change is completely wrong and inappropriate. Such a wrong warning is set off if correct action detection (“A”|A) is followed by incorrect situation checking (“U”|S) when (A, S). Thus we have:

$$P_w(\text{wrong and inappropriate warning}|A, S) = P(\text{“A”}|A)\ P(\text{ “U”}|S) \qquad (8)$$

(c) Unnecessary support either when (NA, S) or (NA, U)

While the driver has no intention to make a lane change (NA), no warning is needed either when vehicles on the adjacent lane are far behind (S) or when a faster vehicle is approaching from behind (U). The right decision and action for the computer is to keep silent either when (NA, S) or (NA, U). However, a warning can be set off unnecessarily under two conditions, as follows. When the true state of the world is (NA, S), the computer would set off an unnecessary warning if its misinterpretation of the driver behavior (“A”|NA) is followed by incorrect situation checking (“U”|S). When (NA, U), on the other hand, an unnecessary warning would be issued if incorrect driver behavior monitoring (“A”|NA) is followed by correct situation checking (“U”|U). Thus we have:

$$P_w(\text{unnecessary warning}|\text{NA, S}) = P(\text{“A”}|\text{NA})\ P(\text{“U”}|\text{S}) \qquad (9)$$

$$P_w(\text{unnecessary warning}|\text{NA, U}) = P(\text{“A”}|\text{NA})\ P(\text{“U”}|\text{U}) \qquad (10)$$

where (9) and (10) cover all the cases for unnecessary warning for (“A”|NA).

4.3 Probabilistic evaluation of the support by action

(a) Accident prevention when (A, U)

When the computer detects the driver’s lane change action (“A”) and it judges that the action does not fit the situation (“U”), it applies a protective control to prohibit the lane change. If the support by action is of the hard protection type, such as one that neglects the driver’s control input for a while, the driver cannot override the computer’s protective control. The conditional probability that such hard protection (hp) can prevent an accident when (A, U) is given by:

$$P_{hp}(\text{accident prevention}|\text{A, U}) = P(\text{“A”}|\text{A})\ P(\text{“U”}|\text{U}) \qquad (11)$$

If the support by action is of the soft protection type, such as one that makes the steering wheel slightly stiffer than usual, the driver could override the protective control by adding more force into the steering wheel. Whether such soft protection (sp) can avoid an accident or not depends on whether the driver abandons his/her original intention for a lane change. Then we have:

$$P_{sp}(\text{accident prevention}|\text{A, U}) = P(\text{“A”}|\text{A})\ P(\text{“U”}|\text{U})\ P(\text{BO}|\text{soft protection}) \qquad (12)$$

where P(BO|soft protection) denotes the conditional probability that the driver breaks off (BO) his/her lane change action upon noticing the soft protection. It is seen, from (7), (11), and (12), that support by action of hard protection type is always the most capable in avoiding an accident.

(b) Wrong and inappropriate support when (A, S)

For either type of protection (hp/sp), the computer applies a wrong protective control when (A, S) if its correct detection of the driver’s lane change action (“A”|A) is followed by incorrect interpretation of the situation (“U”|S). Then we have:

$$P_{hp/sp}(\text{wrong and inappropriate protective control}|A, S) = P(\text{“A”}|A)\ P(\text{ “U”}|S) \qquad (13)$$

The outcome of the ‘wrong protective control’ differs between the hard and soft protections: The driver loses a chance of a proper lane change under the hard protection, while he/she still can make a lane change by applying a stronger force under the soft protection.

(c) Unnecessary support either when (NA, S) or (NA, U)

Even when the driver has no intention to make a lane change (NA), the computer would apply a protective control unnecessarily if its misinterpretation of the driver behavior (“A”|NA) is followed by incorrect situation checking (“U”|S) when (NA, S) or by correct situation checking (“U”|U) when (NA, U). Thus, as with an unnecessary warning, we have:

$$P_{hp/sp}(\text{unnecessary protective control}|NA, S) = P(\text{“A”}|NA)\ P(\text{“U”}|S) \qquad (14)$$

$$P_{hp/sp}(\text{unnecessary protective control}|NA, U) = P(\text{“A”}|NA)\ P(\text{“U”}|U) \qquad (15)$$

where (14) and (15) cover all the cases for unnecessary protective control for (“A”|NA). Note that ‘unnecessary protective control’ does not always bring an explicit inconvenience to the driver: For instance, if the hard/soft protection is applied without any alert, the driver may not notice that the protective control is applied, because the force by the protective control can be felt only when the driver steers the wheel and he/she would not steer the wheel without any intention of making a lane change, which contrasts starkly with the case of the support by warning in which an ‘unnecessary warning’ is annoying.

4.4 Observations

(I) Suppose the computer makes no errors in driver behavior monitoring (P(“A”|A) = P(“NA”|NA) = 1) and in situation checking (P(“U”|U) = P(“S”|S) = 1). Then the *support by action with hard protection* (SBA-HP) provides performance at least as great as either the *support by action with soft protection* (SBA-SP) or the *support by warning* (SBW), because:

(i) none of SBW, SBA-HP, and SBA-SP gives wrong and improper support when the driver’s control action fits the situation; see, equations (8) and (13),

(ii) none of SBW, SBA-HP and SBA-SP provides unnecessary support when the driver is not taking any specific control action; see, equations (9), (10), (14), and (15), and

(iii) SBA-HP never fails to avoid an accident, while SBW or SBA-SP can avoid an accident only when the driver respects the warning or the soft-protection; see equations (7), (11) and (12).

(II) Suppose the computer may make errors in the driver behavior monitoring ($P(\text{“A”}|A) \leq 1$, $P(\text{“NA”}|NA) \leq 1$) and/or in the situation checking ($P(\text{“U”}|U) \leq 1$, $P(\text{“S”}|S) \leq 1$). The SBA-HP’s capability to avoid an accident is still the best among the three. However, even the SBA-HP cannot always avoid an accident, because $P_{hp}(\text{accident prevention}|A, U) \leq 1$.

(III) The computer’s imperfection in driver behavior monitoring and/or situation checking brings problems to each type of support. The first problem is a wrong and inappropriate support that can interfere with the driver when (A, S), which is due to incorrect situation checking $P(\text{“U”}|S)$. Although the probability of a wrong support is the same among SBA-HP, SBA-SP, and SBW, the outcome differs significantly. The SBA-HP’s protective control that activates when (U, A) causes a loss of chance to make a proper manual lane change. In case of SBA-SP, on the other hand, the driver can still make a lane change if he/she puts a stronger force onto the steering wheel. Although the SBW’s wrong warning trying to stop the driver’s proper action may be annoying, the inconvenience would be the least among the three. Note here that it may not be so easy to make $P(\text{“U”}|S)$ smaller. An obvious reason is that engineers’ nature is to be conservative, to allow more false alarms than missed detections. Another reason is that it is almost impossible to satisfy all the drivers with a single alarm threshold to determine whether “U” or “S”. The computer’s judgment (“U” or “S”) may fit to the situation evaluation of some drivers, but may not correspond to the desired criterion of some other drivers with different skill levels or risk perceptions.

(IV) The second problem is an unnecessary support that might be given when the driver has no intention to make a lane change. There are two scenarios in which such an unnecessary support is provided: As equations (9) and (14) show, the first scenario occurs under the condition (NA, S) if incorrect behavior monitoring is coupled with incorrect situation checking (which type of failure would be negligible under the assumption that behavior monitoring and situation checking are statistically independent). The second scenario happens under the condition (NA, U) if incorrect behavior monitoring is coupled by correct situation checking, equations (10) and (15). Although the probability of an unnecessary support is exactly the same among SBA-HP, SBA-SP, and SBW, the outcome is quite different: The SBW’s unnecessary warning may be annoying to the driver. However, unnecessary protective control by SBA-HP or SBA-SP does not necessarily bring clear inconvenience to the driver: As mentioned in 4.3 (c), the driver may not notice at all that the protective control is activated when he/she has no intention to make a lane change, because force feedback by the protective control can be felt only when the driver actually steers the wheel. This argument holds only when the SBA-HP or SBA-SP does not issue any alert when it activates protective control. If an alert sound is given the driver to let him/her know the activation of protective control, either SBA-HP or SBA-SP cannot be free from the problem of unnecessary warnings. Then the key lies in how P("A"|NA) can be made smaller: The probability of incorrect behavior monitoring, P("A"|NA), may not be negligible in case β where the "A" decision or "NA" decision must be made early by inferring the intent of the driver, which contrasts starkly with case α where P("A"|NA) or P("NA"|A) may be small enough; see, section 3.4 (III). One way to make P("A"|NA) smaller is to improve sensing technology for behavior understanding or intent inference technology. Another way might be to allow the computer to conclude the driver's action at the latest time point possible, which means that less time is available for the driver to perceive, understand, and respond to the computer's warning, which may bring disadvantages to the SBW.

5 Concluding remarks

The Convention on Road Traffic (1968) states that "Every driver of a vehicle shall in all circumstances have his vehicle under control so as to be able to exercise due and proper care and to be at all times in a position to perform all maneuvers required of him" (Article 13.1). In reality, drivers sometimes fail to take action that is necessary in the situation (case α) or take action that does not fit the situation (case β). If the human must always be in control, the SBA is not permissible, because it is the machine that analyzes the situation, selects an action, and implements the action. However, as shown in sections 3 and 4, the SBA with machine-initiated trading of authority from human to machine is effective and indispensable for avoiding an accident occurring in cases A and B.

In spite of such theoretical primacy of SBA over SBW, the drivers may not always prefer SBA to SBW. Inagaki, Itoh, & Nagai (2007, 2008) conducted experiments with a driving simulator to compare SBA and SBW in cases α and β from the viewpoints of safety as well as the drivers' acceptance. The experiments were done under the assumption that the computer makes no errors in understanding the traffic situation and the driver behavior, and the specifications for the SBA and SBW were exactly the same as described in sections 3.1 and 4.1 in the present paper. They found that the SBW was not effective enough, compared with the SBA, to prevent an accident when the drivers disregarded warnings for either case α or β. An interesting finding was that the drivers did not accept fully the SBA for case β (more precisely, SBA-HP) in spite of its primacy over the SBW in accident prevention, while they accepted well the SBA for case α. This finding may yield the following conjectures: (i) The human can accept machine-initiated trading of authority if its aim is to take care of what he/she is unable or has failed to do, and (ii) the human may be reluctant to accept the machine-initiated trading of authority if the aim is to prohibit what he/she wants to do. One way to solve (ii) is a coupling of the machine-initiated trading of authority with the human-initiated one so that the human can override the machine's decision when needed by seizing the authority back from the machine. The SBA-SP for case β is such an example. The coupling of the two schemes for authority trading may also be a solution for the problem of improper support by the SBA in case α. Our society is now seeking the design of *human-machine coagency* in which humans and machines are 'equal' partners (Hollnagel and Woods 2005, p. 67) coupled with each other tightly or loosely for shared and cooperative controls (Abbink & Mulder 2010; Flemisch, Heesen, Kelsch, Schindler, Preusche, & Dittrich 2010; Sentouh, Debernard, Popieul, & Vanderhaegen 2010). Both schemes for authority trading (viz., machine-initiated and human-initiated ones) are needed in the realization of the environment in which humans and machines collaborate cooperatively.

References

Abbink, D.A. & Mulder, M. (2010). Motivation for a neuromuscular basis for haptic shared control. *Preprint of the IFAC Human-Machine Systems*. 5 pages in CD-ROM.

Billings, C.E. (1997). *Aviation automation – The search for a human-centered approach*. Mahwah, NJ: Laurence Erlbaum Associates.

Convention on Road Traffic (1968). 1993 version & amendments in 2006, New York, NY: United Nations.

Endsley, M.R. (1996). Automation and situation awareness. *Automation and human performance* (pp.163-181). Mahwah, NJ: Laurence Erlbaum Associates.

Flemisch, F., Heesen, M., Kelsch, J., Schindler, J., Preusche, C., & Dittrich, J. (2010). Shared and cooperative movement control of intelligent technical systems: Sketch of the design space of haptic-multimodal coupling between operator, co-automation, base system and environment. *Preprint of the IFAC Human-Machine Systems*. 9 pages in CD-ROM.

Hollnagel, E. & Woods, D.D. (2005). *Joint cognitive systems: Foundations of cognitive systems engineering*. Boca Raton, FL: CRC Press.

ICAO (1998). *Human factors training manual*. Doc 9683-AN/950.

Inagaki, T. (2003). Adaptive automation: Sharing and trading of control. *Handbook of cognitive task design* (pp. 147-169). Mahwah, NJ: Laurence Erlbaum Associates.

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), pp. 264-272.

Inagaki, T., Itoh, M., & Nagai, Y. (2008). Driver support functions under resource-limited situations. *Journal of Mechanical Systems for Transportation and logistics*, 1(2), pp. 213-222.

Parasuraman, R., Bhari, T., Deaton, J.E., Morrison, J.G., & Barnes, M. (1992). *Theory and design of adaptive automation in aviation systems* (Progress Rep. No. NAWCADWAR-92033-60). Warminster, PA: Naval Air Development Center Aircraft Division.

Rouse, W.B. (1988). Adaptive aiding for human/computer control. *Human Factors*, 30(4), 431-443.

Scallen, S.F. & Hancock, P.A.(2001). Implementing adaptive function allocation. International *Journal of Aviation Psychology*, 11(2), 197-221.

Scerbo, M. W. (1996). Theoretical perspectives on adaptive automation. *Automation and human performance* (pp.37-63). Mahwah, NJ: Laurence Erlbaum Associates.

Sentouh, C., Debernard, S., Popieul, J.C., & Vanderhaegen, F. (2010). Toward a shared lateral control between driver and steering assist controller. *Preprint of the IFAC Human-Machine Systems*. 6 pages in CD-ROM.

Woods, D. (1989). The effects of automation on human's role: Experience from non-aviation industries. In *Flight deck automation: Promises and realities* (NASA CR-10036, pp.61-85).

Zhou, H., Itoh, M., & Inagaki, T. (2009). Eye movement-based inference of truck driver's intent of changing lanes. SICE Journal of Control, Measurement, and System Integration, 2(5), 291-298.