FRAMEWORK FOR UNDERSTANDING THE DRIVER’S TRUST IN AUTOMATION AND ITS IMPLICATIONS ON DRIVER’S DECISION AND BEHAVIOR

HongSeok Cho and R. John Hansman

Project Final Report for Ford-MIT Alliance Project: Investigation of Driver Reactions to Reduced Confidence Alerts in Automated Driving

Report No. ICAT-2018-01
January 2018

MIT International Center for Air Transportation (ICAT)
Department of Aeronautics & Astronautics
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
Executive Summary

Society of Automotive Engineers (SAE) level 2 or 3 automated driving systems have limitations which require appropriate monitoring and/or intervention by the driver in order to obtain the potential benefits of increased safety and efficiency, as well as reduction of driver workload. Inappropriate use of the automated systems due to over-trust or under-trust may lead to safety risks and market value risks. The objective of this research is to understand how designs of the automated driving system influence the driver’s automation use decisions, and to provide design recommendations that would promote appropriate decisions of the driver.

In order to represent the driver’s appropriate automated use decisions, a risk-based framework was developed. Traditional risk management theory was adopted to create a “decision matrix” which represents the driver’s trust in automation with respect to the driver’s willingness to accept perceived risks associated the use of the automation. The matrix presents the risk in terms of two variables: 1) perceived reliability of the automation, and 2) perceived consequence of the automation’s potential unreliable behavior. Each block of the decision matrix represented a level of the perceived risk and an accompanying automation use decision of the driver: 1) the driver may accept the perceived risk and use the automation (the ‘trust’ case), 2) the driver may use the automation with monitoring to mitigate the perceived risk to be considered acceptable (the ‘trust but verify’ case), or 3) the driver may not accept the perceived risk and not use the automation (the ‘do not trust’ case).

Having the decision matrix framework at the core, an overall driver-automation system architecture was developed in order to describe how the driver makes a cognitive assessment of the observable states to evaluate the two decision matrix variables. The architecture includes the driving environment, the vehicle, the automated driving system and various driver feedback channels which together provide a set of observable states to the driver. The driver perceives and comprehends the observables and makes a projection of the future driving condition. That is, based on this projection and the driver’s understanding of the automation, the driver evaluates the automation’s reliability for a projected driving condition. Concurrently, the driver predicts potentially unreliable behaviors of the automation to evaluate the consequence in terms of a recoverable margin to avoid hazardous consequences. The decisions based on these evaluations create plans to use, to use but monitor, or not to use the automation. Note that the decisions to use over a given planning horizon, may or may not include plans for monitoring the driving situation and the behavior of the automation in order to update the risk evaluation and to detect and recover from unreliable behaviors of the automation.

The architecture also includes a learning process of the driver aside from the real-time evaluations. Once driver experiences a performance of the automation, the driver associates the observed performance (reliable or unreliable) with a set of observed states to evaluate the automation’s conditional reliability. In principle, the assessment is compared with past experience in order to create or modify heuristic-based knowledge of the automation. This understanding is used to create the driver’s mental model of the automation which is used in to evaluate the decision matrix variables.

The framework’s decomposition of the driver’s decision into the two evaluations suggests that a potential inappropriate decision can be a result of the inaccurate evaluations of either the automation reliability or the consequence or both. The architecture can be used to demonstrate if the incorrect automation reliability evaluation is due to the driver’s incorrect mental model of the automation, the driver’s incorrect
projection of the future state, or both. Incorrect association of the observed reliability with a set of observables, loss of understanding due to infrequent observables may cause the driver to have misleading learning process and result in an incorrect mental model of the automation. Unexpected changes in the driving environment, imperceptible observable states, inadequate sampling of observable information due to an incorrect monitoring plan or distractions may cause the driver to have an incorrect projection of the future states. Uncertainty of future behaviors of the automation, improper recovery plans, and improper confidence in the manual recovery may lead to incorrect evaluation of the consequence.

In order to avoid potential inaccurate reliability and consequence evaluations of the driver, the architecture will suggest and enable analysis of various mitigation methods with which the manufacturers can design systems to support the driver to make appropriate evaluations of the decision matrix variables. For instance, alerting systems such as the Reduced-confidence alerting system (Tijerina et al., 2016) can provide a notification to the driver of unreliable driving conditions which may but lead to an automation failure. The alert has a potential to support the driver in detecting less salient observables (e.g. degraded roadway), adequately sampling necessary information and correctly associating observable states (e.g. worn lane line markings) with the unreliable behavior of the automation (e.g. automation loss-of-tracking), thereby promoting both appropriate learning of system limitations as well as fostering appropriate automation use decisions (e.g. close monitoring in unreliable driving conditions). However, if the alerting system is not accurate, the low positive predictive value of the alert may impact the driver’s mental model of the system and the driver may ignore the alert. Display of the automation’s criteria may support the driver to better interpret the alerts and the automation’s behavior. The framework will also be developed to promote more appropriate monitoring of the system and driving conditions and guide understanding of the driver’s learning process.

Using the framework, the mitigation methods for promoting appropriate automation use decisions can be developed and evaluated. The framework can both facilitate the design of the simulator trials and serve as a basis to determine the effectiveness of the methods including displays or behaviors built into the automation. The outcome of the research expects to identify opportunities for Ford Motor Company to enrich the design of the potential automated driving systems with effective driver supports that promote appropriate and informed decisions of the driver regarding the use of the automation.
Outline

1. Research Motivation and Objectives ................................................................. 5

2. A Risk-based Framework to Represent the Driver’s Automation Use Decisions .................. 6
   2-1. Automation Use Decision Matrix
   2-2. Driver’s Appropriate Automation Use Decisions

3. Overall Driver-Automation Systems Architecture ............................................. 11

4. Real-time Cognitive Processes that Influence the Driver’s Decision ............................. 12
   4-1. Evaluating the Decision Matrix Variables: Automation Reliability and Consequence
   4-2. Driver’s Mental Model and Situational Awareness
   4-3. Monitoring
   4-4. Planning and Implementation

5. Learning: The Driver’s Development of the Mental Model through Past Experience ............. 18
   5-1. Learning Through Indirect Experiences
   5-2. Learning Through Direct Experiences

6. Factors that May Induce Inappropriate Automation Use Decisions ............................. 21
   6-1. Incorrect Reliability Evaluation
   6-2. Incorrect Consequence Evaluation

7. Potential Mitigations to Support the Driver to Avoid Inappropriate Automation Use Decisions 25
   7-1. Automation Performance
   7-2. Limit Automation Use
   7-3. Alerting Systems
   7-4. Information Feedback
   7-5. Automation Interface
   7-6. Manuals and Training

8. Conclusions and Recommended Future Research ............................................. 30

References ........................................................................................................... 31
1. Research Motivation and Objectives

SAE Level 2 or 3 automated driving systems have limitations with human factors implications (SAE, 2016). In a Level 2 system, the social contract is that the driver must continue to monitor the driving situation. The driver is also expected to intervene manually if necessary, whether or not a Take-Over-Request (TOR) is issued. This expectation is based on the notion that the Level 2 system may sometimes ‘not know what it does not know.’ In a Level 3 system, the automated driving system technology has evolved to the point where the driver no longer needs to monitor the driving situation but must respond to a TOR if issued. The L3 concept is that a TOR will be provided in sufficient time for a typical driver to respond adequately. If the driver does not respond, the L3 system is intended to be able to achieve a safe state (e.g. stopping) but perhaps with limitations (e.g., stopped in the travel lane rather than parked on the roadside). An L3 system’s automation will deal with routine and emergency conditions and events within its operational envelope so the driver need not monitor. However, the driver’s perception of system reliability and consequences might still influence how quickly they resume manual driving and this is a topic of great interest to the automated driving research and development community.

In order to obtain potential benefits (Kalra, 2017) (i.e. increased safety and efficiency, reduction of driver workload) of the automated driving systems, the driver needs to use the system appropriately. Inappropriate use of the automated systems may lead to various risks. For example, the driver may decide to use the system without appropriate monitoring due to overreliance, resulting in safety risks. Conversely, the driver may decide not to use the system when the system is available due to under-trust and this results in loss of the potential safety, efficiency, and workload reduction benefits. It is also possible that the driver may intervene or interrupt the automation when the automation would provide performance superior to that of the driver, e.g., when the automation is executing a maneuver. Also, when the driver under-trusts the vehicle, the driver may not appreciate the performance of the system that they have purchased or considered purchasing which results market value risks to the manufacturer.

Past studies on operators’ use of automation systems in aviation field suggest that designs of the system influence the operator’s decisions regarding the use of the automation. In order to design automated driving systems that promote the driver’s appropriate use of the automation, it is critical to understand how the drivers make automation use decisions and how various system designs (e.g. alerting systems, information displays and automation performance set-points and requirements) may influence those decisions.

The objective of this research is to develop an architecture and associated methodology to understand how designs of the automated driving system influence the driver’s automation use decisions, and utilize the methodology to provide design recommendations that would promote appropriate decisions of the driver.

The research is intended to identify potential ways for Ford Motor Company to enrich the design of automation systems with effective driver supports. The architecture and modeling framework reported on here may be used as the starting point of a systematic guide to the development of automated systems features that promote adequate driver understanding and trust. These features may include guidance on appropriate content for information displays (e.g. situation displays, predictor displays, event or post-event notifications); guidance and verification criteria for the content for a digital assistant to explain automation operation, capabilities, and limitations in context-specific ways (e.g., to subsequently evaluate
if the content is modifying behavior as intended). The architecture may guide the development of automated driving system alerts and warnings, along with their system settings (e.g., positive predictive value for a lowered confidence alert). The architecture may identify control intervention strategies that will mitigate the likelihood that a driver manually intervenes in an automated driving maneuver in a sub-optimal manner. Perhaps the greatest opportunity will be in the architecture’s ability to formally capture processes by which drivers learn system capabilities and limitations over time. This might result in recommendations on how to facilitate that learning and retention of real-world driving cues in the environment-automation interaction that provide information on system reliability and consequences of an automation’s unreliable behavior.

2. A Risk-based Framework to Represent the Driver’s Automation Use Decisions

The research literature has shown that human’s use of automation is influenced by his/her trust in the automation system (Lee & See, 2004). Trust has been defined in past studies as the human’s willingness to accept risk based on his/her belief that the automation will exhibit reliable behavior under uncertainty of a given situation (Cho et al, 2015). This belief is based on the human’s cognitive assessment of past experience with the automation. Incorporating this idea, this study represented the driver’s decision to use of the automated driving systems in terms of the driver’s acceptance of risk associated with the use of the automation for a given driving condition based on their experience with the system and its reliability.

2-1. Automation Use Decision Matrix

In order to represent the driver’s decisions effectively, the concept of a risk matrix from the traditional risk management theory (ISO/DIS, 2009) has been adopted. The risk matrix is a formal representation of the risk associated with a system or a procedure.

This concept is used in various fields such as aerospace both for program risk management and to determine certification criteria for acceptable risk (FAA, 2009). The risk matrix represents the risk in terms of two variables: likelihood of an adverse event and consequence of the event. Based on the two variables, the matrix is used to determine the acceptability of the associated risk. Higher risk would be considered unacceptable and lower risk would be considered acceptable. Moderate levels of risk may be considered acceptable when the risk can be mitigated through various methods.

This idea has been adopted in this research as a “decision matrix” which conceptually describes the driver’s decisions regarding the use of the automation as shown in Figure 1. Analogous to the risk matrix, the decision matrix represent driver’s perceived risk associated with use of the automation in terms of two variables: the driver’s perceived reliability of the automation (analogous to likelihood of an adverse event in the risk matrix) and the driver’s estimated consequence of the automation’s potential unreliable behavior (analogous to consequence severity in the risk matrix). The decision matrix is used as a surrogate to describe both, the appropriate decision to use the automation and the driver’s cognitive processes to make the automation use decisions. Note also in Figure 1 that the SAE levels of automation have been superimposed on the Decision matrix in certain cells. For example, it is likely that only highly automated L4 and L5 systems will be unconditionally reliable and may operate without any need for intervention. Indeed, this concept is inherent in L4 & L5 concept vehicles such as the Google concept that operate without any driver controls at all (Gannes, 2014).
Figure 1: Driver’s Automation Use Decision Matrix and the Appropriate (Rational) Use Decisions for SAE Automation Levels 2-5

The traditional risk matrix concept was not linearly mapped into the decision matrix for the driver. The variables of the risk matrix in a domain such as aviation or nuclear power plant operation (IAEA, 2001) are based on rigorous numerical calculation of the likelihood of a system’s failure and the system’s failure consequence is estimated in terms of the severity (e.g. levels of property damage or human injury). In contrast, the variables of the decision matrix in Figure 1 are based on real-time subjective evaluation by the driver that is based on limited information and relatively shorter time frames for evaluation.

Therefore, in order to fit our research objective of understanding the driver’s decisions, the preliminary definitions and categories of the decision matrix variables have been defined below. This categorization aimed to adequately represent the driver’s automation use decisions with a minimum resolution. The resolution of the categories may be higher or lower for different individuals. However, drivers with low resolution (simpler matrix) may have insufficient sophistication for appropriate decisions.

Perceived Reliability

The driver’s perceived reliability is defined as the driver’s perception of the reliability of the automation system to behave reliably during a projected future driving condition or planning horizon that might be identified in time, distance, or both. In normal driving, this planning horizon may be as short as 1 second ahead for curve negotiation (e.g., Land and Lee, 1994), or several seconds ahead for tactical maneuvers like intersection approach and negotiation (Lunenfeld and Alexander, 1990), or even longer periods for strategic levels of driving like route following. This subjective assessment is based on the driver’s understanding of the automation’s capabilities and limitations, which may vary between individuals depending on their personal experience with the automation.
The levels of the perceived automation reliability are notionally categorized in Table 1. The levels are decomposed into unconditional and conditional assessment of the automation’s reliability.

For unconditionally reliable systems, the driver expects the system to behave reliably regardless of the driving condition. The driver does not assess the driving condition because he/she expects that the system can handle every possible condition that may arise during the entire or defined segment of the trip. In practice, however, the drivers may spontaneously periodically visual sample of the driving situation to ensure all is well (Blommer, Curry, Kochhar, Swaminathan, Talamonti, and Tijerina, 2017). Once the driver has learned or developed the impression of ‘unconditional reliability’, the driver will only reassess the driving condition when the system requests the driver to do so or when a noticeable unreliable behavior of the automation occurs. Similarly, for unconditionally unreliable, the driver does not expect the system to behave reliably regardless of the driving condition and presumably would elect to drive manually.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unconditionally</td>
<td>The automation will behave reliably regardless of the driving condition.</td>
</tr>
<tr>
<td>Reliable</td>
<td>(i.e., ‘it’s perfect’)</td>
</tr>
<tr>
<td>Conditionally</td>
<td>The automation will behave reliably for the projected driving condition,</td>
</tr>
<tr>
<td>Reliable</td>
<td>but may not be reliable for other conditions.</td>
</tr>
<tr>
<td>Conditionally</td>
<td>It is uncertain if the automation will behave reliably</td>
</tr>
<tr>
<td>Less Reliable</td>
<td>for the projected driving condition.</td>
</tr>
<tr>
<td>Conditionally</td>
<td>The automation will behave unreliably for the projected driving condition,</td>
</tr>
<tr>
<td>Unreliable</td>
<td>but could be reliable for other conditions.</td>
</tr>
<tr>
<td>Unconditionally</td>
<td>The automation will behave unreliably regardless of the driving condition.</td>
</tr>
<tr>
<td>Unreliable</td>
<td></td>
</tr>
</tbody>
</table>

Outside the extremes of perfect reliability or perfect lack thereof, the driver may assess the driving condition to assess the reliability of the automation with respect to it. This “conditional reliability” of the automation due to the fundamental limitations of the SAE level 2 or level 3 automated driving systems which performance may degrade depending on various driving conditions. Drivers may or may not possess this understanding of condition-dependent capabilities and limitations of the automation depending on their experience and/or prior knowledge.

The levels of condition-dependent perceived reliability are categorized into three levels. 1) For conditionally reliable, the driver expects the system to behave reliably for the projected driving condition. The duration of that condition depends on the extent of the driver’s projection of the reliable condition. 2) For conditionally less reliable, the driver is uncertain whether the automation will behave reliably or unreliably for the projected driving condition. 3) For conditionally unreliable, the driver does not expect the automation to behave reliably for the projected driving condition.

The scale of trust in the automation may vary between individuals from binary (e.g. reliable and unreliable) to higher resolution, based on the driver’s levels of sophistication of the understanding and individual differences. However, if a driver has lower resolution (i.e. binary), the driver would make only simple decisions that would be inadequate to experience the automation in various driving conditions. For
example, if a driver only perceives an automation to be either reliable or unreliable, the driver is may not choose to use the automation with appropriate monitoring to recover from an unreliable behavior. Also, the driver may not be able to learn about driving situations

Perceived Consequence

The consequence severity in the traditional risk matrix is often categorized in terms of severity of the injuries to humans and damage to property. However, it is less likely that the driver will estimate the level of physical damages during real-time assessment and projection of the future event. Also, because for ordinary driver’s the use of automation is ‘preference-based delegation’ of the driving task to the automation instead of ‘dependence-based delegation’ (Castelfranchi, 2009) the research framework hypothesizes that the driver would not accept hazardous consequence from the use of the automation. Therefore in this research, the perceived consequence is defined as the driver’s evaluation of whether a potential hazardous consequence due to unreliable behaviors of the automation can be safely recovered. When the automation starts to behave unreliably, there is a temporal or spatial leeway until the unreliable behavior results a hazardous consequence (e.g. crash). This research framework called this leeway a ‘remaining margin for recovery’ or a ‘recovery margin’. In order for the driver to safely recover from a hazardous consequence, the driver needs to detect the unreliable behavior and implement an effective recovery action within the remaining margin. This remaining margin is further decomposed and categorized into four levels as shown in Table 2 below.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Need for Intervention</td>
<td>No recovery action will be needed to avoid consequential event.</td>
</tr>
<tr>
<td>Recoverable with Sufficient Remaining Margin</td>
<td>There will be significant margin for the driver to avoid hazardous consequence after the automation starts to behave unreliably.</td>
</tr>
<tr>
<td>Recoverable with Little or No Remaining Margin</td>
<td>There will be little or no margin for the driver to avoid hazardous consequence after the automation behaves unreliably. (Immediate recovery action will be needed)</td>
</tr>
<tr>
<td>Unrecoverable (Lacking Remaining Margin)</td>
<td>There will be no enough margin for the driver to avoid hazardous consequence after the automation starts to behave unreliably.</td>
</tr>
</tbody>
</table>

This remaining margin could be infinite, sufficient, just enough or not enough. Infinite margin indicates that there is no necessary action needed for the driver to safely recover. Sufficiently remaining margin provides large leeway for the driver to detect and recover. Little or no remaining margin provides tight leeway for the driver to avoid a hazardous consequence thus require immediate actions. When the remaining margin is insufficient for a safe recovery, the only way for the driver to avoid a hazardous consequence is to takeover to manual control before the automation starts to behave unreliably. Relatively little is known about how a human operator develops estimates of safety margin, how accurate they might be, and how they might be modified with training, feedback or experience. For example, it
has been known for many years that drivers ‘over-drive’ their headlights at night, i.e., drive faster than the headlight illumination distance allows for effective braking to a stop if necessary (Leibowitz, 1986). It is an interesting question if a field-of-safe travel display might alter a driver’s behavior in such situations. Analogues to such a situation in automated driving may be investigated in future work.

2-2. Driver’s Appropriate Automation Use Decisions

Based on the assessed risk as reflected in the decision matrix, the driver will make decisions which will create plans regarding the use of the automation. The resulting decisions are illustrated with different colors for the decision matrix cells in Figure 1. If the driver’s assessments have correctly evaluated the two decision matrix variables, the colors would represent appropriate (rational) decisions of the driver for the evaluated level of risk. Sections of the matrix partitioned with bold boundaries in Figure 1 indicate appropriate automation use decisions for different SAE automation levels.

Green represents a driver’s decision to use the automation without monitoring of the driving condition, automation state, and the interaction between the two. The driver determines that there is no risk associated with the use of the automation because the driver assesses that the automation will behave reliably regardless of the future driving condition and therefore the driver does not need to manually intervene. This decision and behavior would be consistent with for SAE Levels 4-5.

The driver could also rationally decide to use the automation without monitoring if the driver trusts that the automation will conditional behave reliably, and when the condition deteriorates, the automation would provide a notification to the driver with a sufficient time for the driver safely takeover to manual control. The decision would be consistent with for SAE level 3 where it is assumed that the automation will handle emergency conditions and ask the driver to take over only in relatively limited cases with adequate time to respond.

Yellow and orange represent various forms of a driver’s decision to use the automation with driver monitoring which would mitigate the risk associated with the automation use. The two colors (yellow and orange) notionally represent different levels of monitoring behavior of the driver. Yellow represents a driver monitoring driving condition. Orange is where the driver closely monitors the automation’s behavior to execute immediate actions if necessary. This also assumes that the driver processes to create recovery plans. The driver will monitor the automation’s behavior with the thresholds to determine an unreliable behavior. When the behavior is determined unreliable, the driver will implement a planned recovery action. These levels of monitoring are discussed in more detail later in this report.

Red represents driver’s decision not to use the automation. The driver would rationally choose not to use the automation when the driver determines that the risk associated with automation use cannot be mitigated by driver’s monitoring of the driving condition and the automation’s behavior. These decisions would be consistent with for SAE Level 2.
3. Overall Driver-Automation Systems Architecture

In order to model how the drivers make the automation use decisions, a driver-automation systems architecture has been developed as shown in Figure 2. The architecture is an information flow model to describe the cognitive processes of the driver.

![Figure 2: Driver-Automation Systems Architecture](image)

The architecture the decision matrix as a core component by which the driver creates plans and implements those plans regarding the use of the automation. Because the decision matrix represents the driver’s automation use decisions in terms of two input variables, two evaluation processes (i.e. evaluation of automation reliability and evaluation of consequence of automation’s unreliable behavior) are represented to provide the two input variables for the decision matrix.

The driver conducts the reliability and consequence evaluations based on the future dynamics that are projected through his/her mental model of the driving environment and automation systems comprehended by the driver’s situational awareness process.

The driver’s mental model is influenced by the learning process where the expected and actual observed automation behaviors together with observed driving condition create a set of experiences. The set of experiences is cognitively processed to create, maintain or modify the driver’s knowledge (understanding) of the automation, thus updating the driver’s mental model of the automation system.

The details of these cognitive process and the learning process are discussed in the following section. The understanding of the processes by which the driver develops and utilizes the mental model and the knowledge of the systems provides a basis for identifying factors that may result inappropriate decisions and designing mitigations to support the driver to avoid the undesired decisions.
The driving environment, the vehicle and various systems which the driver interact with are shown is the left side of the overall architecture. The automation system observes the driving environment through sensors and provides control input the vehicle. The control inputs generate the vehicle states (e.g. velocity and acceleration) through the vehicle dynamics, and the vehicle states interact with the various environmental states such as other vehicles, bicycles, pedestrians and/or road infrastructures.

These interactions create observables that could be perceived by the driver. The driver could perceive the external environment and the vehicle states through windshields and mirrors. The driver could also perceive the vehicle maneuver such as the orientation and accelerations through the vestibular system. The driver could also perceive the vehicle and automation status through driver feedback from the auditory or visual displays such as an alert, a dashboard display or an automation interface console.

4. Real-time Cognitive Processes that Influence the Driver’s Decision

In order for the driver to make automation use decisions, the driver engages in real-time perception, comprehension, and projection of the available information (observables). The driver’s awareness is used to make real-time evaluation of two variables used in the decision matrix (perceived automation reliability and perceived recovery margin). If the driver decides to monitor, the driver monitors the driving condition and the automation behavior in real-time in order to update his/her evaluations and/or implement planned actions. This section discusses in detail the cognitive processes that influence the driver’s automation use decision.

![Figure 3: Evaluation of the Decision Matrix Variables: Perceived Reliability and Perceived Consequence](image-url)
4-1. Evaluating the Decision Matrix Variables: Automation Reliability and Consequence

The reliability and consequence evaluations for the automation use decision are shown in Figure 3. The cognitive information used for those evaluations and the resulting decision matrix variables will be discussed below.

Automation Reliability Evaluation

For all but unconditional trusted or untrusted situations, the driver evaluates the reliability of the automation for current observed and future predicted driving condition. This evaluation results the driver’s perceived reliability of the automation which influences the driver’s automation use decision in the decision matrix. In order to make the evaluation, driver needs the projected states from the mental model and his/her situational awareness.

The driver’s mental model includes the driver’s understanding of the automation’s performance associated with various “contexts” and we called this the “conditional reliability of the automation”. The context consists of driving environmental states (e.g. weather condition, road condition, and lane curvature), traffic situation (e.g. locations and maneuvers of other road users) or displayed information/alerts (e.g. reduced-confidence alert, display of automation’s awareness states via a real-time situation awareness display that indicates what the automation sees and planning to do next) or some combination of these. Based on the past experience, the driver has an associated level of automation reliability with a context. For example, the driver may have associated an unreliable behavior of the automation with a context of a reduced-confidence alert state. This development of an understanding is discussed in more detail later in this report on “learning”.

Once driver has a set of projected states from the situational awareness process, the driver determines if the set of states matches one of the contexts from his/her understanding of the automation. If the driver finds a context that matches the observed and projected states, the level of reliability associated the context will be the driver’s perceived automation reliability for the projected driving condition. This process is consistent with the recognition-primed decision (RPD) concept, where a person understand a situation in terms of relevant cues and associated expectancies (Klein, 1993). It is unclear what would the driver do when he/she cannot find a matching context from his/her past-knowledge (i.e. encountering a situation for the first time). Since it is uncertain whether the automation would behave reliably or not, the driver may choose to continue to monitor the actual behavior of the automation to generate a new context, or the driver may refer to prior-knowledge gain from training or other similar system and decide to not use the automation due to expected or unexpected behavior of the automation. It is a research question to figure out what would be appropriate decisions and behaviors of the driver for unforeseen driving conditions, and how to support them to make the appropriate decisions.

Consequence Evaluation

Through the consequence evaluation, the driver evaluates whether a potentially hazardous consequence due to unreliable behavior of the automation can be safely recovered by the driver. In order to make the evaluation, the driver may use the mental model to mentally simulate a potential event for which the automation behaves unreliably. For the mental simulation, the driver needs: 1) the expected unreliable behaviors of the automation and an understanding of how that behavior would manifest itself, 2) the
future projection of the driving situations and how they would change based on the unreliable automation behavior, and 3) a recovery plan based on what appears to be appropriate and adequate.

Through the mental simulation, the driver evaluates if the driver will be able to safely recover from an expected unreliable behavior and prevent a potentially hazardous consequence (e.g. a crash). Driver determines of the recovery margin is sufficient, just enough or insufficient for his/her plan for a recovery action sequence. This evaluation results in the driver’s perceived recovery margin for potential unreliable behaviors of the automation which, in turn, influences the driver’s automation use decision in the decision matrix.

4-2. Driver’s Mental Model and Situation Awareness

A mental model is a cognitive representation of the driver’s awareness and knowledge of a process (Rouse & Morris 1985; Seppelt & Lee 2015). In this context, the model includes representation of the driving environment, the automation system, the alerting system and the driver’s current plans. The driver’s use of his/her mental model to perceive available observables, comprehend them to obtain the driver’s situational awareness (Endsley, 1995), and to project the comprehended states to make future state predictions. This awareness of the situation are supported by the driver’s understanding. The understanding includes the environment, traffic situation, automation system, alerting system and the driver’s current plans. The understanding of the automation system is used to make reliability evaluation based on the projected driving condition. Also, the driver uses the mental model to cognitively simulate potential hazardous events caused by unreliable behavior of the automation in order to evaluate the remaining margin for the driver’s recovery action.

Since the driver’s use of the mental model to evaluate the decision matrix variables have been discussed above, this section is focused on how the mental model is used to make a projection of the future states through the situational awareness process.

![Figure 4: Drive’s Mental Model and the Situation Awareness Process](image)

The driver forms his/her mental model with perceived information and his/her knowledge developed from past experience. The mental model is used to 1) comprehend the perceived information, 2) project the current comprehension to predict future states, and 3) use these predictions to search and sample information from the available observables in order to maintain or update his/her mental model. The accuracy of the prediction of the mental model is critical for the driver’s appropriate decision because the driver evaluates the decision matrix variables (perceived reliability and perceived potential consequence)
based on the driver’s perception of current states and prediction of the future states of the automation, vehicle, and driving environment.

Recent studies have shown that these predictions are generated in a hierarchical manner. Engström et al (2017) adopted a concept called “predictive processing” in order to represent the human cognitive process as it might unfold in a driving context. They hypothesize that the human constantly generates predictions and compares the predictions with what is actually observed. When a prediction error is detected, the human will either modify the prediction or take action to maintain the original prediction.

Within the predictive processing framework, the predictions are generated from what is called a ‘hierarchical generative model’, which is equivalent to the mental model in our framework. The framework hypothesizes that the mental model generates predictions at different levels of abstraction, and a higher level prediction results in or generates lower level predictions.

Engström and his colleagues applied this framework to the driving task by incorporating Michon’s (1985) terminology. According to Michon, the driving task has three levels of abstraction: 1) the maneuver level, relating to the momentary control of the vehicle (e.g., lane keeping, speed maintenance, and inter-vehicle separation); 2) the tactical level, relating to the current driving condition and maneuvers (e.g., lane change, curve negotiation, intersection negotiation); and 3) the strategic level, relating the general goals of driving (e.g., route selection; wayfinding; driving style to trade off fuel efficiency, safety, travel time, etc.). This concept of hierarchically decomposed levels of prediction is useful for our study to represent how the different-level-predictions are used to evaluate the two decision matrix variables, and how the drivers monitor the actual driving situation in order to compare those predictions with the actual observed states.

Based on the observed driving condition and/or his knowledge (e.g. general knowledge of traffic system or familiarity with the road), the driver will make a prediction about the future driving condition over some planning horizon that may range from a fraction of a second to several seconds (or even longer for strategic planning like route following). For example, if a driver observes a degraded road condition ahead, the driver will predict that the vehicle may be driving on faded lane markings. If a driver is familiar with the road, e.g. by exposure during an everyday commute, the driver will expect that he/she will approach a sharp curve shortly. This type of driving condition predictions reflects tactical level predictions according to predictive processing framework.

Based on these higher (tactical) level predictions about the driving condition, maneuver level predictions will be generated by the driver. The driver will make predictions regarding how the automation will or should behave in the predicted driving condition. For example, for the expected upcoming sharp curve, the driver expects the vehicle to slow down at a certain distance before the curve, steering wheel to turn towards the turning road direction. These tactical and maneuver level predictions will be used during the driver’s monitoring of the driving condition and the automation’s behavior as described below.

4-3. Monitoring

Driver monitoring is a key activity for the driver to mitigate the assessed risk associated with the automation use in SAE level 2 automation so that the risk can be considered acceptable. In our research, the monitoring process serves two purposes: 1) to determine a deviation of the future state projections
that would influence the driver’s evaluation of the decision matrix variables (i.e. where you are in the matrix), and 2) to determine an unreliable behavior of the automation in order for a recovery action sequence to be initiated.

Since the automation’s reliability is associated with the driving condition that is constantly changing, the driver’s perception of the automation reliability needs to be updated accordingly. As discussed in the predictive processing framework, the driver will generate tactical level predictions of the driving condition. Monitoring processes will determine if there is a change in the driving condition prediction by comparing between the previously predicted driving conditional states and the current predicted states. Any deviation from the driving condition projection will be used for a re-evaluation in order to update the perceived automation reliability and the perceived consequence.

Depending on the complexity or uncertainty of the driving condition, the driver’s confidence on the prediction of the driving condition will vary. This will influence the extent of the prediction that the current predicted states will stay constant. This varying confidence will result monitoring plans with different sampling rate. For example, straight rural highway, low traffic level without intersections and will have relatively constant driving condition compared to the urban roads with high traffic volume and intersection with crosswalks.

When the driver determines that the automation will be reliable for a tactical level prediction of the driving condition over a driver-defined planning horizon or time period, this will generate maneuver level predictions that the automation will behave reliably. The driver may choose not to monitor the automation down to the maneuver level. Conversely, if the perceived automation reliability is determined uncertain for a predicted driving condition, the driver must generate maneuver level predictions for potential unreliable behavior of the automation. The driver needs to monitor down to the maneuver level in order to compare the predicted automation behavior with the actual behavior of the automation. The sampling rate for this maneuver level monitoring task will most likely be higher compared to the tactical level monitoring for the driving condition, thus creating multiple levels of monitoring. For the maneuver level monitoring, the driver’s projection of the unreliable behavior becomes a threshold to determine if the unreliable behavior has actually started. If the behavior has reached the threshold, the driver will start the planned recovery action sequence to avoid a potentially hazardous consequence.

4-4. Planning and Implementation

Based on the driver’s automation use decision, the driver generates plans regarding the use of the automation.

Automation Use Plan

First, based on the level of assessed risk from the decision matrix, the driver generates plans to use the automation. This plan may simply be to initiate an action sequence or continue the current use of the automation if already engaged. The ‘do not use’ plan when the automation is not engaged is simply a plan to continue to manually drive. If an unacceptable risk is assessed (‘red’ in the decision matrix) while using the automation, the driver will come up with a takeover plan and implement the plan to manually control the vehicle. If the automation is able to detect an unacceptable risk condition, it would issue a TOR and the driver would be expected to respond to the request. If the automation detects reduced-confidence
conditions, it might issue a reduced-confidence or ‘heads’-up’ alert to prompt the driver to monitor even though the system is still able to function to a reduced degree. The driver may preempt to resume manual driving even before there is an automation degraded performance. This takeover before an actual observation of an unreliable behavior is preemption. The driver would preempt to resume manual control when the driver perceives an unreliable driving condition. Preemption might also arise when the driver perceives that a predicted hazardous consequence is unrecoverable due to insufficient remaining margin and chooses to avoid the risk.

*Monitoring Plan*

If the driver has chosen to use the automation with monitoring, the driver will generate monitoring plans to sample the necessary information from the available observable states. Based on the extent of the future projection with confidence that the reliable driving condition would stay constant for some planning horizon, the sampling rate for updating the situational awareness will be determined. Also based on the driver’s knowledge of the observable states that would influence the automation reliability, the monitoring plan will guide the driver to look for specific observables that are related the conditional reliability of the automation.

*Recovery Plan*

When the driver perceives that the automation has become conditionally less reliable and no plenty recovery margin is expected, the driver will prepare for potential unreliable behavior of the automation by generating a recovery plan. The recovery plan includes 1) a threshold for the unreliable behavior determination, and 2) a recovery action sequence to intervene and resume manual control when the unreliable behavior is determined. Depending on the expected unreliable behavior of the automation a recovery plan will vary. Uncertainty in the unreliable behavior projection may result multiple failure modes which require multiple recovery plans. The uncertainty may lead the driver to over-estimate the consequence and preempt to takeover due to anxiety. If the decision is premature, this will prohibit the driver from monitoring and be witnessing an automation’s reliable behavior. The driver may inappropriately associate the driving condition with an ‘imagined’ unreliable behavior and not use the automation, which will be discussed in more detail in the later.

*Risk Reduction Plan*

If a high risk is assessed, the driver may plan and conduct risk reduction strategies, instead of passively reacting to the assessed risk by turning the automation ON or OFF. If the automation system permits the driver to modify the automation settings (e.g. a headway distance to a lead vehicle, automated lane change) while the automation is engaged, the driver may modify those settings to reduce the assessed risk. The driver may create a plan to transition to more reliable driving condition. For example, if degraded lane markings are observed and the driver observes that a lane adjacent to the travel lane has a better road condition, the driver may simply conduct an automated lane change to transition to the adjacent lane which would increase the perceived reliability of the automation. The driver may also create a plan to increase the margin for recovery. For example, the driver may adjust the automation setting to increase the headway distance to a lead vehicle or reduce the set speed so that the driver has more distance to slow down to avoid a crash. If the remaining margin for recovery seems tight, the driver may have his/her hands on the steering wheel and/or have a foot on the brake/accelerator pedal depending on the recovery action plan. This too might provide an increase in the recovery margin which can critical to avoid a crash.
5. Learning: The Driver’s Development of the Mental Model through Past Experience

Learning is a process by which the driver develops his/her knowledge, skills, expectancies, and attitudes. In this modeling architecture, the learning is focused on the driver’s development and evolution of his/her understanding of the automation’s capabilities and limitations. As discussed above, the driver uses this understanding of the automation in order to make an automation reliability evaluation for a projected driving condition. The understanding guides the driver to monitor for specific observable states that may influence the automation reliability. Also, the understanding based on learning may support the driver to predict unreliable behavior of the automation which will be used during the recovery margin evaluation. The research hypothesizes that the driver develops and evolves this understanding over time through experience.

SAE level 2 or 3 systems, by their nature, are not fully robust and therefore the system has performance limitations that are dependent on various driving conditions. As the driver experiences this conditional dependency of the automation’s performance, the driver will understand that there is ‘conditional reliability’ of the automation performance. The process of developing this understanding of automation’s conditional reliability through experience is defined as a “learning process” which will be discussed further in the following section of the report.

![Figure 5: Learning Process](image)

The learning process is modeled in Figure 5 above. In the developed architecture, the experiences which influence the learning process was categorized into two kinds: 1) indirect experience, and 2) direct experience.

5-1. Learning Through Indirect Experiences

Indirect experience is defined as the driver being exposed to information through not directly based on interactions with the automation. Various indirect experiences include reading the owner’s manual,
participating in training sessions, watching YouTube videos on various automated driving systems, or hearing a friend’s experience on his/her purchased vehicle, or simple hearsay. It should be noted that the indirect experiences may vary widely in the accuracy and completeness of what is learned.

These indirect experiences create initial heuristics regarding the automation’s capabilities and limitations prior to the direct experience. The initial heuristics will be confirmed or modified as the driver directly experiences the automation. It is desirable to have an adequate initial understanding of the automation so that drivers are not vulnerable during their early experience with the automation. For example, if the driver’s initial understanding that the automation would slow down for a sharp curve is incorrect, the driver may not be able to adequately detect and recover from the unexpected behavior of the automation. However, a real-world experience of learning the operation of Adaptive Cruise Control (ACC) indicates that drivers closely observe the new technology and learn over time how it does and does not operate (Larsson, 2012)

Reading the manual will provide comprehensive understanding of the automation; however studies have shown that the drivers may not thoroughly remember the vast information written in the manual, or some drivers may not even read the manual. Manufacturers may design interaction manual or provide training programs prior to the purchase in order to effectively provide necessary information.

Reading the owner’s manual and Company-prepared videos can provide comprehensive understanding of the automation. However, drivers often do not read their owner’s manuals (Melenbacher, Wogalter, and Laughery, 2002). It is also the case that dealership staff sometimes do not accurately convey how active safety technology works (Abraham, McAlnuty, Mehler, and Reimer, 2017). Even when correct information is conveyed, drivers may not comprehend or remember all aspects of system operation, limitations, and capabilities. This leaves an opportunity for manufacturers to design digital assistants for use in the vehicle to explain how to use the automation, interpret displays, and comprehend driving conditional states associated with an event (i.e., post-event explanation).

5-2. Learning Through Direct Experiences

Direct experience is defined as the driver’s direct interactions with the automation in a real driving situation. The information obtained by the driver through direct experience consists of 1) expected automation behavior, 2) observed actual automation behavior and 3) observed (and remembered) states prior to and during the actual automation behavior observation.

As discussed earlier, the mental model contains the driver’s understanding of the automation’s performance as the “conditional reliability”, which consists of various “contexts” and associated levels of reliability of the automation performance. Learning through direct experience is a process in which the information obtained from a recent experience is assessed using the past-understanding in order to create, confirm or modify the driver’s mental model. Lee and See (2004) asserted the importance of this context-dependent understanding, “Appropriate trust depends on the operator’s understanding of how the context affects the capabilities of the automation.” The process developing this understanding is discussed in detail below.

Based on the driver’s projection of the future state, the driver determines if the projected set of states has a matching context in the mental model. If the driver determines that there is one, the associated
level of reliability becomes the expected behavior of the automation. If the set of projected states does not match previously generated context, the driver may generate a new context which does not have an expected behavior yet. Each context consists of a set of observable states including driving environmental states (e.g. weather condition, road condition, and lane curvature), traffic situation (e.g. locations and maneuvers of other road users) or displayed information/alerts (e.g. reduced-confidence alert, display of automation’s awareness states and future plans) or some combination of these.

When the driver has observed an actual behavior of the automation, the driver determines if the automation has behaved as expected for the matching context (if the context exists in the mental model). If the automation behaved reliably as expected, this observation will confirm the driver’s perceived reliability of the automation. If the automation behaves unexpectedly, the driver may try to rationalize the recent experience by trying to find a difference between the observed (and remembered) states and the context used to make the expectation for the recent experience. If the driver cannot find the difference (i.e. unable to rationalize), the driver’s may modify the associated reliability for the context that the automation will behave reliably (i.e. make a transition to lower levels of perceived reliability).

Conversely, the driver may choose to use the automation even if the driver expects an unreliable behavior of the automation. This driver’s decision can arise if the driver wants to test his/her understanding of the automation and the recovery margin is sufficient to safety test the automation. If the automation unexpectedly behaves reliably, and if the driver cannot find the difference in the context, the driver’s may modify the associated reliability that the automation will behave unreliably (i.e. make a transition to higher levels of perceived reliability).

The driver may be uncertain if the automation would behave reliably or unreliably for a context (i.e. the driver is considering both likelihoods of reliable and unreliable behavior). This may be due to not enough accumulated experience in order to determine whether the automation would behave reliably or not, or the driver may have observed both the reliable and the unreliable behavior of the automation for that context. Accumulation of consecutive reliable or unreliable behaviors of the automation for the same context may modify the associated uncertainty into a certain perceived reliability, conditionally reliable or conditionally unreliable, respectively. However this “tipping point” may vary depending on the nature of the contexts. Certain contexts may be clearer to make a reliability judgement than others. For example, an observation of automation unable to detect and respond to a stop sign can be quickly be associated with unreliable performance of the automation, whereas a level of degraded weather condition that would cause an unreliable automation behavior is harder to determine and the automation’s performance during the degraded weather condition may fluctuate depending on various other states (e.g. curved road, visibility of lane markings). Even though it is difficult for estimate the tipping point, based on the decision matrix framework, the driver’s automation use decisions and monitoring behaviors will indication changes in the driver’s perceived automation reliability for a driving condition.

The driver cannot, and is not expected to generate every context that would completely determine reliable unreliable behavior of the automation. The driver may choose to stay at the uncertain level of perceived reliability for certain contexts and choose to monitor the automation’s behavior to detect and recover from a potential unreliable behavior. The key to appropriate learning is a generation of adequate contexts and associating appropriate levels of reliability to the generated contexts. It is shown in a recent study (Beggiato et al, 2013, 2015) that the direct experience can modify incorrect prior understandings or
create understandings which were not obtained during prior indirect experiences. Therefore supporting the driver to develop an accurate understanding of the automation during their direct experience is crucial.

6. Factors that May Induce Inappropriate Automation Use Decisions

The information flow represented in the developed architecture provides an opportunity to trace back from the decision matrix in order to identify where and when, in the driver’s cognitive process, an incorrect assessment, incorrect understanding or an absence of necessary information may occur that would result in an inappropriate decision.

The decision matrix framework has decomposed the mechanism by which the driver makes the automation use decision into two cognitive processes, the driver’s evaluation of automation’s reliability, and the driver’s evaluation of the remaining margin for safe recovery from potential unreliable behavior of the automation. Correspondingly, the factors that may induce inappropriate driver’s decisions are the factors that cause 1) inaccurate reliability evaluation (e.g. over-trust or under-trust), and/or 2) inaccurate consequence evaluation (e.g. over-estimation or under-estimation of the recovery margin in time or space). Various factors that may induce the incorrect evaluations are identified in Table 3 below and discussed in detail in this section.

6-1. Incorrect Reliability Evaluation

In order to evaluate the automation’s reliability, the driver project future states based on his/her situational awareness to find a matching context and an associated level of automation reliability based on his/her understanding of the automaton (mental model). Therefore the incorrect reliability evaluation may arise from 1) the driver’s incorrect projection of the future states and/or 2) the driver’s incorrect understanding of the automation (i.e. incorrect mental model).

The driver projects future states based on the perceived and comprehended observable states. If the driver does not perceive and/or comprehend necessary observables that influence the automation’s reliability, the projected states used to make the reliability evaluation may be inadequate. The driver may not perceive observables simply because they are not salient. For example, the driver may not notice degraded lane markings. This can be also due to driver’s incomplete understanding of the automation. The driver would actively search for context that is associated with unreliable behavior of the automation. However if the driver does not know degrade lane markings may lead to loss-of-tracking, the driver may not notice the indistinct state.

The driver may also have incorrectly projected states if the driver inadequately samples new observables to maintain the accurate state awareness. Driver updates his/her situational awareness through monitoring, and the sampling rate of the monitoring process depends on the extent of the driver’s projection of the reliable condition. If the driver is over-confident for the duration of his/her projection of the reliable condition, the driver may sample less frequently. And if there is an unexpected change in the driving condition that alters the automation reliability, the driver may not detect the change promptly. Also, even if the driver had a monitoring plan with an adequate sampling rate, the driver may be distracted,
or the driver may be fatigued which would cause unintentional decrease of the sampling frequency and result in an incorrect situational awareness.

Even if the driver has correctly projected the future states, the driver’s incorrect or incomplete understanding of the automation (conditional reliability) may lead to an incorrect reliability evaluation. The incorrect or incomplete understanding may develop from an inappropriate learning. Potential cases of the inappropriate learning are discussed below.

While learning, the driver may associate a level of automation reliability (reliable or unreliable) with an incorrect context (i.e. observable states). If a driver associates a context, which can actually be reliably handled by the automation, with an unreliable behavior the driver may become ‘superstitious’ and falsely expect an unreliable behavior for a reliable context and decide to not use the automation inappropriately. Furthermore, the driver would not witness the actual reliable behavior of the automation for the context since the automation is not used. This incorrect understanding is due to the driver’s incorrect assessment of an unreliable behavior of the automation. Since there are various observables that may have caused the unreliable behavior, the driver may make an incorrect association especially if the affecting observable states are not salient.

An inappropriate learning may also result from the driver’s association of a certain reliability level to an inadequately broad context. For example, when a driver observes an unreliable behavior while driving in a local road, the driver may associate the unreliable behavior to the entire local roads instead of more detailed contexts such as a sharp curve or an intersection with a stop sign. Conversely, the driver may also falsely associate a reliable behavior to a broader context without further assessments. For example, after experiencing a reliable behavior of an automation on the same portion of a highway for a few trips, the driver may consider the automation is reliable for the entire highway. This association of reliable or unreliable behaviors of an automation to an inadequately broad context may cause the driver to inappropriately use the automation for unsafe conditions without appropriate monitoring or to inappropriately not use the automation for reliable conditions, respectively. Lee and See (2004) called this inappropriate association a low “specificity” of trust, which can result an incorrect calibration of trust.

If the driver inappropriately decided to use the automation without monitoring, the driver would only perceive the actual behavior of the automation when it behaves unreliably in a noticeable manner. Observable states prior to the unreliable behavior will not be observed. Driving condition which caused the unreliable behavior may or may not be observed by the driver depending on the persistency of the conditional state. Likewise, if the driver does monitor due to his/her expectation that the automation would provide a takeover command with a sufficient margin to recover, the driver would only perceive the behavior of the automation if a command is given prior to the unreliable behavior. If the automation has failed to detect its own unreliable behavior, the driver would only witness the outcome of the behavior when the behavior results in a consequence. It is likely that the driver cannot make a rational assessment of the failure event due to lack of information to support the assessment. Worst case, the driver may falsely associate the unreliable behavior with an irrelevant context.

An inappropriate learning may also result from the difference between the automation’s criteria and that of the driver. The recovery plan includes driver’s plan on when to intervene. This timing to initiate a recovery action sequence would be set by the driver’s threshold for determining the unreliable behavior of the automation. This threshold may vary between drivers depending on their driving habit or an
Table 3: Driver’s Inappropriate Automation Use Decisions and Influencing Factors

<table>
<thead>
<tr>
<th>Incorrect Evaluations</th>
<th>Incorrect Information/understanding</th>
<th>Cause of the incorrect information/understanding</th>
<th>Resulting Inappropriate Automation Use Decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Incorrect future state projection</td>
<td>Unobserved necessary observable states</td>
<td>Indistinct observable states</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Inadequate (top-down) information acquisition</td>
</tr>
<tr>
<td></td>
<td>Inadequate monitoring (insufficient sampling rate)</td>
<td>Incorrect projection of the duration of constant reliable condition</td>
<td>Inappropriate Use</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inappropriate automation use decisions</td>
<td>Driver distracted or fatigued</td>
</tr>
<tr>
<td>Incorrect reliability evaluation</td>
<td>Incorrect mental model</td>
<td>Incorrect association</td>
<td>Incorrect association between a context and a level of reliability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unobserved automation performance</td>
<td>inappropriate generalization of a level of reliability with a broad Context</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not monitored states due to inadequate monitoring</td>
<td>Driver’s preemption due to a discrepancy in criteria between an automation and a driver</td>
</tr>
<tr>
<td></td>
<td>Lost (forgotten) knowledge/understanding</td>
<td>Rarely occurring context</td>
<td>Inappropriate Use</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Indistinct observable states</td>
<td>Inadequate mental model</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inadequate mental model</td>
<td>Inappropriate Use</td>
</tr>
<tr>
<td></td>
<td>Incorrect hazardous state projection</td>
<td>Unobserved hazardous states</td>
<td>Inadequate mental model</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inadequate monitoring</td>
<td>Inappropriate automation use decisions</td>
</tr>
<tr>
<td></td>
<td>Incorrect projection of the unreliable behavior</td>
<td>Incorrect projection of the unreliable behavior</td>
<td>Inadequate mental model</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Over-estimation of the unreliable behavior</td>
<td>High uncertainty (anxiety)</td>
</tr>
<tr>
<td></td>
<td>Lack of confidence in the manual recovery</td>
<td>Individual differences in experience/skills/strategies</td>
<td>Inappropriate Disuse (inappropriate preemption)</td>
</tr>
<tr>
<td></td>
<td>Over-confident in the manual recovery</td>
<td>Indistinct observable states</td>
<td>Inappropriate Disuse (inappropriate preemption)</td>
</tr>
</tbody>
</table>
acceptance of risk. For example, when stopping for a slower lead vehicle, the timing to initiate braking may vary between individuals. If an automation’s control algorithm differs from a driver’s standard, this discrepancy may lead to driver’s early initiation of the recovery action prior to the automation’s action. Since the driver was not able to observe the actual response of the automation, this may result in an inappropriate learning by developing an understanding that the automation is not reliable for the context: a ‘superstitious’ understanding of the driver.

The driver may have a correct knowledge of a certain context-dependent limitation of the automation either from reading the owner’s manual (indirect experience) or an experience of observing the unreliable behavior. However the context may be rarely occurring and if the driver does not experience the context and the associated unreliable behavior for a lengthy period, the knowledge may be forgotten. This incomplete mental model may lead the driver to miss-detect affecting observable states while monitoring, or even falsely determine that the context would not affect the reliability of the automation. For example, depending on driver’s nominal driving pattern the driver may rarely experience a sun-glare that may affect the visibility of the sensor and result a loss-of-tracking or a miss-detection of an obstacle. So when the context is presented, the driver may not remember its impact on the automation reliability even if the driver had previously read the information from the owner’s manual.

6-2. Incorrect Consequence Evaluation

In the developed framework, the driver evaluates a potential consequence of a possible unreliable automation behavior in terms of a sufficiency of the remaining margin between an unreliable behavior and a potentially hazardous consequence. The driver first projects a potential hazardous event based on interaction between an expected unreliable behavior of the automation and projected environmental states. The driver considers potential recovery plans (i.e. when to intervene and how to intervene) and determines if the remaining margin is sufficient for the recovery action. Therefore, the driver may inappropriately evaluate the consequence if the driver: 1) incorrectly project a potentially hazardous event, or 2) incorrectly or inappropriately plan a recovery action sequence.

Based on the driver’s mental model (understanding of the automation, past observations), the driver may project a potential unreliable behavior of the automation for a projected driving condition. For example, when a driver observes or projects a degrade road condition for a straight road, the driver may expect the automation to slowly drift to either side of the road. If the degraded lane markings are presented on a curved road, the driver may expect a sudden lane departure of the vehicle. Based on the expected unreliable behavior the driver would estimate the remaining margin (e.g. time or space between a start of a lane departure to a crash with the guardrail on the roadside).

However, actual unreliable behavior of the automation would be hard to project unless the automation’s plan is apparent to the driver. Both an under-estimation (e.g. expecting a slow drift when the automation is actually planning an abrupt turn) and an over-estimation (vise-versa) of the unreliable behavior would lead to an incorrect evaluation of the consequence. An under-estimated unreliable behavior may lead to an over-estimation of the remaining margin and may result in an unrecoverable hazardous consequence. Conversely, an overly estimated unreliable behavior may lead the driver to preempt to takeover due to perceived unrecoverable consequence (i.e. insufficient remaining margin). Also, when a driver is uncertain about the future behavior of the automation, the driver may imagine various failure modes which can
increase the anxiety and, again, lead to an over-estimation of the unreliable behavior. If the automation is, in fact, able to handle the driving condition, this over-estimation of the unreliable behavior (i.e. under-estimation of the remaining margin) would prohibit the driver from experiencing the actual performance of the automation and loss an appropriate learning opportunity.

In order to evaluate whether the estimated remaining margin is sufficient or not, the driver assesses if the driver can detect the unreliable behavior and execute a potential recovery action sequence safely within the estimated margin. This assessment may be incorrect if the driver is falsely confident on his/her capability to manual recovery. If the driver over-confident in his manual recovery plan, the driver may not be able to recover since the recovery action actually required more time or space than expected. Conversely, if the driver is not sure that he/she can safely recover, when there is sufficient margin to execute the recovery plan, the driver will preempt and lose the opportunity to observe the actual behavior of the automation. This improper confidence in manual recovery may arise due to individual differences in their experience, skills, and strategies.

7. Potential Mitigations to Support the Driver to Avoid Inappropriate Automation Use Decisions

The risk-based framework provides an approach to evaluate potential mitigations that may support the driver to avoid the inappropriate automation use decisions identified in the previous section. A set of potential mitigations are summarized in Table 4 below. The mitigations include human-automation system approach to determine required automation performance, defining operational limits of automation use, alerting systems, information feedback and automation interfaces, and development of training material.

<table>
<thead>
<tr>
<th>Potential Mitigations</th>
<th>Mitigation Opportunity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automation Performance</td>
<td>• Determine minimum performance requirements to provide sufficient margin under assumed failure conditions</td>
</tr>
<tr>
<td>Limit Automation Use</td>
<td>• Limit use of automation for driving conditions with insufficient margin</td>
</tr>
<tr>
<td>Alerting Systems</td>
<td>• Support driver’s reliability evaluation when the reliability is reduced</td>
</tr>
<tr>
<td></td>
<td>• Support driver’s monitoring</td>
</tr>
<tr>
<td>Information Feedback</td>
<td>• Display automation’s decision criteria to support driver’s comprehension of the automation</td>
</tr>
<tr>
<td>Automation Interface</td>
<td>• Provide means to adjust margin</td>
</tr>
<tr>
<td>Training</td>
<td>• Support the development of accurate mental model of the automation</td>
</tr>
</tbody>
</table>
This section of the report will discuss the potential mitigations identified based on the architecture. The potential mitigations will be further investigated and evaluated for their effectiveness during the next phase of this research.

7-1. Automation Performance

The human-automation system includes components, which often have limitations or uncertain performance requiring human monitoring and intervention. In the SAE automation levels, those systems would be designated level 2 or level 3 systems. System components which may be limiting are indicated with yellow in Figure 6. The components include; sensors, data processor, logic and algorithm of the automation system. The risk-based framework provides a natural structure for system level risk analysis of potential automation system, which may be used to determine the minimum performance requirements for the system components to provide sufficient margin for the driver to recover from potential risk events under failure conditions to meet an assumed level of safety.

![Figure 6: System Design Components that would Influence the Decisions and Behaviors of the Driver](image-url)
In order to conduct a risk analysis, it is first necessary to identify potential failure modes or limitations of each of the system component and associated driving conditions. For example, if an automation system uses a camera to visually detect lane markings in order to control the steering angle, the visibility of the lane markings may be impacted by faded lane markings or weather (e.g. road covered with snow or reduced visibility by fog). Based on the identified failure modes or limitations of the system components, for each of those failures, the architecture can be used to identify the ability of the human to recognize the failures and recover with sufficient margin.

For a particular system, the architecture can be used to identify the observable states (e.g. vehicle feedback, hazards in the environment, information feedback) that driver may perceive in order to identify unreliable behavior of the automation prior to and during the identified risk event. Based on the expected observables and potential hazardous consequence, the cognitive processes of the driver can be modeled in the architecture and used to determine 1) if the driver has adequate information to make correct evaluations of the perceived reliability and the perceived consequence, and 2) if the risk event provides sufficient margin for the driver to detect and recover from a hazardous consequence.

The decision matrix framework can be used to determine if there is adequate monitoring potential for the driver to safely recover within a defined recovery margin. If, for a certain risk event, the observables do not provide sufficient information for the driver to accurately evaluate the risk, the driver may be able to appropriately monitor the system and intervene.

This risk analysis provides an opportunity to assess if it is reasonable to expect for a driver to accurately evaluate the automation reliability and the consequence for a set of risk event for a particular system, and to safely recover from hazardous consequences. The result of the analysis can be used determine minimum performance requirement for the system components or evaluate system level trade-offs in component choice.

7-2. Limit Automation Use

If the risk analysis indicates that a sufficient margin for driver monitoring and recovery may not be possible due to technology limitations, excessive costs or challenging situations an alternative is to restrict or limit automation use to those conditions where acceptable risk levels can be maintained.

This conditional restriction of the automation use is similar to defining the operational design domain (ODD) for level 3 to 5 automation system (NHTSA, 2016). The recent recommendation by the NTSB (2017) emphasized the importance of the ODD concept for Level 2 automation as it is for the Levels 3–5. The report recommended a development of a method to verify that vehicles equipped with Level 2 vehicle automation systems incorporate system safeguards that limit the use of automated vehicle control systems to those conditions for which they were designed. The decision matrix framework may provide the method for determining when to operationally limit the automation use.

7-3. Alerting Systems

Well designed alerting systems have the potential to improve the driver’s automation use decision making by supporting the automation reliability assessment and consequence of failure evaluation. In addition,
alerting systems can help assure that the driver maintains the appropriate level of monitoring for the situation. Based on the cognitive processes modeled in the architecture, the alerting system may 1) support the driver’s reliability evaluation by notifying changes in the driving condition that may reduce the reliability of the automation, 2) support the driver’s monitoring by notifying when the driver does not appear to be adequately monitoring, and 3) support the driver’s consequence evaluation by notifying the presence of hazardous factors in the driving environment that may influence the consequence.

The notification of the driving conditions with reduced reliability of the automation assumes that the automation can detect those factors that may influence the automation reliability. The goal of the alert would be to shift the driver’s perceived automation reliability to a lower level in the decision matrix framework. By perceiving the lower reliability level, the driver would more closely monitor the automation’s behavior to promptly detect and recover from unreliable automation behavior, or preempt to resume manual control.

If the system can detect that the driver is not adequately monitoring, an alert can be issued to the driver and/or the automation can be turned off automatically, if the system determines that the current level of monitoring isn’t sufficient. The goal of the alert is to ask the driver to attend the driving situation so that the driver may correctly evaluate the reliability and the consequence of failure. Since the system cannot directly monitor the driver’s cognitive process, various methods have been or are being considered by the manufacturers as a mean to indirectly detect inattention of the driver. Tesla uses the driver’s hands on the steering wheel as a cue to assess the driver’s engagement (NTSB, 2017). Whereas Volvo has announced plans to use eye-tracking technology, and GM has introduced the driver attention monitoring system based on the camera-based detection of the driver’s head-rotation. The effectiveness of the various methods on correctly detecting the driver’s attention is still an on-going area of research, and can be considered in risk analysis supported by the framework.

If the system can detect hazardous factors that may influence the potential consequence of an unreliable behavior of the automation, an alert or supporting information may support the driver’s perception of important cues for evaluating the consequence.

In order for those potential alerts to be effective, the driver’s mental model must be consistent with the alerting system to trigger the appropriate response. However, the driver’s mental model is derived from their experience with the automation and learning process modeled in the architecture provides an opportunity to understand how the driver’s response to different alerting systems may evolve based on the driver’s experience with the alerts. For example, if an alert notifying a reduce reliability condition accurately predicts unreliable behaviors of the automation (i.e. high positive predictive value), the driver will associate the reduced-confidence alert with a lower reliability state. However, for certain conditional states, it may be difficult to have a high positive predictive value of the alert. If driver repeatedly experiences reliable automation behavior while the alert is active (i.e. low positive predictive value), then the driver’s mental model will evolve to distrust and ignore the alert.

Similarly, if an alerting system notifying inadequate monitoring or engagement of the driver incorrectly issues an alert when the driver is actually monitoring, the experience may lead to the driver to lose trust in the alerting system and ignore the alert or consider the alert intrusive. Ineffective design of the alerting system may not induce expected response from the driver. The NTSB report (2017) on the Tesla fatal accident found that the way that the Tesla Autopilot system monitored and responded to the driver’s interaction with the steering wheel was not an effective method of ensuring driver engagement.
7-4. Information Feedback

Display of information such as automation’s internal logic criteria may support the driver’s understanding and mental model of the automation system. This can influence the driver’s reliability and the consequence of failure evaluation processes.

If the information provided through a display informs the driver what the automation has detected through sensors (e.g. the display may represent lane boundaries, a lead vehicle or a pedestrian detected by the automation), the driver may better understand the automation’s behavior and their evaluation of the reliability of the automation.

Providing such information along with an alert has a potential to mitigate the loss of trust in the alerting system by supporting the driver’s understanding of the cause of an alert so that the driver can understand if the alert was issued on a reasonable basis. For example, when a reduced-confidence alert is issued without any other information regarding why the alert was issued, the driver may not comprehend the presence of conditional states which reduce the automation’s reliability, especially if the conditional states are not salient. As a result, the driver may lose trust in the alert if the alert does not lead to an actual unreliable behavior of the automation. Conversely, if supporting information is provided (e.g. provide an information to the driver that the alert was issued based on the detection of degraded/faded lane markings), the driver may understand that the automation is using the lane markings as a cue to control the vehicle. With the understanding, the driver can consider the alert was issued reasonably even if the alert does not lead to an actual failure of the automation.

Furthermore, information may also promote appropriate learning of the driver by helping the driver to detect less salient relevant observable states and to correctly associate the observable states with a level of automation reliability. It may also remind the driver of rarely occurring driving conditions that affect the automation’s reliability so that the driver may respond appropriately and prevent the necessary knowledge or understanding from being lost (forgotten). The information might be provided by means of post-event notification that explains why an alert was issued or a why the automation behaved the way it did.

7-5. Automation Interface

As discussed in the driver’s planning process discussion, in some situations, the driver may reduce risk by adjusting their margins and proximity to threats. In some cases, this can be done through the automation’s control parameters. For example, if the driver perceives that the consequence of failure is high due to the recovery margin being insufficient due to the close distance from a lead vehicle, the driver may adjust the headway distance between the lead vehicle or slow down by adjusting cruise speed. These strategies to reduce the risk are only available if the automation interface allows the driver to do so.

The decision matrix framework provides a basis to identify strategies for the driver to manage the assessed risk by adjusting automation parameters. Providing adequate options or event suggestions to promote the increase in the recovery margin may increase the possibility of safe recovery by the driver.

Furthermore, the automation may be designed to execute the risk mitigating strategies automatically (e.g. slowing down for a curve or increasing a headway distance for less reliable conditions). These strategies
will be more effective if information that explains why the automation executed the maneuver so that the driver can rationalize the automation’s behavior.

7-6. Manuals and Training

One of the obvious approaches to supporting the development the driver’s mental model is to provide opportunities for appropriate indirect learning such as manuals and training. The understanding of driver’s learning process from the architecture may support development of training materials such as owner’s manual, online user manual content; car dealer training for demonstration; instructional videos, that would support the driver’s development of an accurate mental model of the automation.

The learning process suggests that the driver obtains apriori knowledge through the indirect learning and reasoning from analogous systems. This prior-knowledge is then updated by the driver’s direct experience with the automation.

One of the challenges in developing training material is the broad range of driver experience and sophistication in the driver’s mental models. For example, some people may have very simplistic mental models and think of the automation as a black box or poor observation skills resulting in a weak understanding of the conditional reliability of the automation. Others may have a detailed understanding of different components of the system or strong observational skills with a consequently more sophisticated mental model of the system. The goal of designing manuals and training materials is to determine and inform the spectrum of mental models but the limitations of the human-automation system performance will normally be driven by the lowest level mental model of the system.

Careful consideration of the apriori mental model and the training material can be useful in the design, validation, and verification of the automation system. Vakil and Hansman (1998) proposed the “operator directed process” as an attempt to redefine the design process to incorporate and specify the expected operator mental model of the system behavior as a verifiable design requirement. One version of this approach is to write the training material before the system is developed and use it as part of the automation system validation requirements as opposed to the traditional approach of developing the automation system and then attempting to developing training material which represents the behavior of the complex automation system. Vakil suggested that “by constraining the design process to implementations which can be used effectively by operators at the basis of a robust mental representation, it is expected that less error prone automation systems will be developed.”

8. Conclusions and Recommended Future Research

An automation use decision matrix was developed as a risk-based framework to represent the driver’s appropriate automated use decisions. The matrix decomposes the driver’s automation use decision into two variables: 1) perceived reliability of the automation, and 2) perceived consequence of the automation’s potentially unreliable behavior. An overall driver-automation system architecture was developed in order to describe how the driver evaluates those two decision matrix variables. The architecture includes driver’s cognitive processes of how observables are perceived and understood to make projections. Learning processes are described that suggest how the driver develops the
understanding of the automation through experience and how that understanding influences the driver’s monitoring, evaluating and planning for the use of the automation.

The framework represented potentially inappropriate decisions of the driver as a result of the inaccurate evaluations of either the automation reliability or the possible consequences, or both. The framework suggests various methods (e.g. reduced-confidence alerting system, training, information feedback or digital assistance) which manufacturers can design to support the driver in learning of the automation’s limitations and thus making appropriate evaluations of the decision variables.

The framework sets the foundation for a next step of this research program. Based on the developed architecture, potential designs of these methods to promote appropriate automation use decisions will be developed for test and evaluation in the next phase of this research. The framework will be used to support the design of the experimental or simulator trials and also serve as a basis to determine the effectiveness of the methods. The goal of the next phase of the research is to identify HMI concepts, automation system settings, automation logic, or some combination of these for Ford Motor Company to enrich the design of automation systems with effective driver supports that promote appropriate and informed decisions of the driver regarding the use of the potential automated driving systems.

References


Rouse, W. B., and N. M. Morris. 1985. On looking into the black box: Prospects and limits in the search for mental models (DTIC #AD-A159080). Atlanta, GA: Center for Man-Machine Systems Research, Georgia Institute of Technology


