Towards Safety Analysis of Interactions Between Human Users and Automated Driving Systems

Fredrik Warg, Stig Ursing, Martin Kaalhus, Richard Wiik

Towards Safety Analysis of Interactions Between Human Users and Automated Driving Systems

Fredrik Warg
RISE Research Institutes of Sweden
Borås, Sweden
fredrik.warg@ri.se

Stig Ursing, Martin Kaalhus and Richard Wiik
Semcon Sweden AB
Göteborg, Sweden
{stig.ursing, martin.kaalhus, richard.wiik}@semcon.com

***Abstract*—One of the major challenges of designing automated driving systems (ADS) is showing that they are safe. This includes safety analysis of interactions between humans and the ADS, a multi-disciplinary task involving functional safety and human factors expertise. In this paper, we lay the foundation for a safety analysis method for these interactions, which builds upon combining human factors knowledge with known techniques from the functional safety domain.**

The aim of the proposed method is finding safety issues in proposed HMI protocols. It combines constructing interaction sequences between human and ADS as a variant of sequence diagrams, and use these sequences as input to a cause-consequence analysis with the purpose of finding potential interaction faults that may lead to dangerous failures. Based on a this analysis, the HMI design can be improved to reduce safety risks, and the analysis results can also be used as part of the ADS safety case.

***Index Terms*—Functional safety, human factors, human error, human performance, HMI, automated driving systems, safety.**

I. Introduction

Automated driving systems (ADS) are seen as having many potential benefits. One of the benefits most often mentioned as a motivating factor for introducing such systems is increased road safety, since common understanding is that driver mistakes are a contributing factor in most serious accidents. However, one of the major challenges for designing an ADS is indeed to show that it is sufficiently safe. Removing the human from the loop is a double-edged sword in this regard. A main strength of a human driver (HD) is handling variability and the shortcomings of technical systems, e.g. handle risky but rare situations, detect and react to mechanical failures [1], or compensate for inappropriate behaviour by other human drivers on the road. In the functional safety standard for road vehicles, ISO 26262 [2], one of the parameters used to estimate the necessary risk reduction in the electrical/electronic (E/E) system of a vehicle function is to which extent the HD or other road users can mitigate a potential hazard. However when driving with full automation the passengers can in general not be expected to mitigate hazards. On the contrary, in addition to the ADS taking over all tasks normally handled by the HD, interactions between the ADS and humans also need to be analyzed to sufficiently reduce the risk of human-machine interaction (HMI) hazards. A difficulty with including interaction analysis in the traditional safety case for an E/E-based function is the multi-disciplinary nature of the work; functional safety experts typically lack the required human factors expertise and vice versa. In this paper we outline a method for safety analysis of interactions with the aim of bridging the gap between the two disciplines. Another interesting method taking both fields into account which is inspired by FMECA [3] is described in [4]. In contrast our method is more focused on analysis of interaction sequences and aimed at the automotive domain.

Different kinds of interaction hazards may be relevant depending on the design choices and automation level of the ADS. For instance, an interaction hazard with road users outside the ADS equipped vehicle could be that the ADS makes sudden and unpredictable maneuvers causing dangerous situations for drivers of other vehicles or cyclists. For interaction between a system and the driver of the vehicle an example from lane keeping assist functions is the experience that too strong corrective action from the automation can startle the driver causing dangerous counter-actions. Hence the subject of interactions between humans and an ADS is wide-ranging [5]. In this paper we focus on one specific class of interactions; transitions of control of the dynamic driving task (DDT) between an HD and an ADS of level 4 according to the SAE taxonomy [6]. We assume a vehicle where control can be transferred from HD to ADS and back while the vehicle is moving, i.e. both a human and an ADS act as drivers during a trip. This can be relevant for e.g. a highway pilot ADS feature which can be enabled when the vehicle is on highways, but is not available on other roads. We believe the described method can be generalized to analyze some of the other types of interactions as well, but leave this as future work.

For transition protocols there are at least three types of hazards to consider: *mode confusion* is a situation where the Human User (HU) and the ADS do not share belief of who is performing the DDT; *unfair transition* is a hazard where either ADS or human is forced to take control in a situation where they are not prepared and able to drive; and *stuck in transition* means either part is unsuccessful in completing a transition for such an extended period of time that the driving capability is impaired. These hazards, together with implementation suggestions and a safety analysis, were previously described by Johansson et al. [7]. However, this work lack an

This research has been supported by Vinnova via the project ESPLANADE (ref 2016-04268).

explicit method to aid in finding the possible human interaction failure modes, or in other words in which ways the human-machine interaction could go awry. This paper proposes a safety analysis methodology which systematically identifies interaction failures between the human and the machine during transition of control of the DDT.

The process consists of the steps: (1) propose a transition protocol; (2) create the interaction sequence with HU and ADS as two communicating entities through the HMI, considering the possible combinations of time intervals; (3) perform cause-consequence analysis (CCA) by constructing cause-consequence diagrams (CCD) based on the interaction sequences, and for each failed event on the CCD perform a fault tree analysis (FTA) considering a model of human behavior; and lastly (4) perform a risk assessment for identified potential faults and improve the HMI design if the residual risk is considered unacceptable. We also suggest that the results of the analysis can be used as a part of the argument for safety of the ADS, and thus used in the ADS safety case.

We illustrate the method using an example transition protocol. It should be noted that we do not propose specific HMI solutions, the example is solely presented to illustrate the method. We also focus on non-malicious interaction faults, i.e. interactions that would fall under the subject of cybersecurity is not in our scope. Finally, we would like to stress that, of course, both risk assessment and verification and validation of solutions need to be backed up by data from real-world tests, but this is also a subject beyond the scope of this paper.

II. Background & Related Work

A. Terminology

The taxonomy in [8] is well established in the dependable systems community, and at least the definitions of fault, error and failure illustrated in Fig. 1 are also used with similar definitions in most functional safety standards; this is also how we use the terms in this paper.

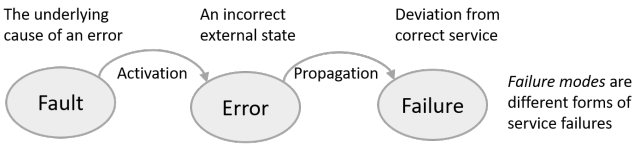

Fig. 1. Fault-Error-Failure sequence from [8].

The results of human actions in the form of faults due to human developer or operator involvement are accounted for according to the fault classification in Fig. 2, and collectively referred to as human-made faults in [8]. These are, however, more seldom used in a functional safety or dependable systems context, as the interaction with human users is often considered out of the scope.

In a human factors context, the terms human error (e.g. [9]), use error (e.g. [10]) and human performance (e.g. [11]) are often used, even if the use of the word error in itself is contested [12]. A united definition of what constitutes a human error, how it should be viewed and how human error should be

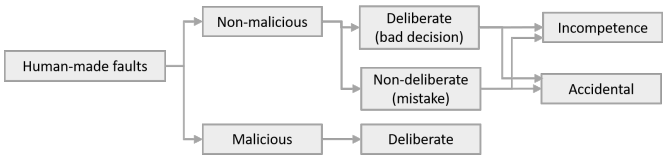

Fig. 2. Taxonomy of human-made faults from [8].

avoided is much debated [9], [13], [14]. In many cases, when an accident occurs, if no explanation can be found that refers to the technical part of the system, the inferred fault "human error" is used [15]. Not until one takes a closer look does it become apparent that the term 'human error' is insufficient to identify the underlying causes of faults by human action [12], and improve the combined performance of the human and technical system.

In the safety analysis method proposed in this paper, human error is not viewed as the cause, but the effect (or symptom) of issues affecting human action performance [16], e.g. interface design, cognitive workload, and biases [17]. Due to this, we prefer not to use the taxonomy of [8] for human error, since this taxonomy derives the cause of human error as either accidental or incompetence, which seems insufficient as a basis for finding these underlying issues. Rasmussen [18] and Reason [19] are examples that goes into detail of how humans perceive, comprehend, project, decide and act in different ways. Their respective classifications of human error are combined in Fig. 3. Rasmussen's skill-, rule-, and knowledge-based framework describes three types of processes requiring different levels of cognitive workload, from the mostly automated skill-based behavior, to the use of learned rules and procedures, to the knowledge-based behaviour required in new situations. Reasons framework, on the other hand, focuses on different types of errors; slips - which are attention errors, lapses - which are memory errors, or mistakes - which are decision-making errors. Violations are a class separate from human error, and occurs when the human is intentionally doing something wrong. We will use concepts from [12], [18], [19] that allows the analysis to look behind the term human error. In order to create a method to systematically investigate all relevant aspects of the human behavior for each step in an HMI protocol, we found the situation awareness and decision making model [17] described in detail in the next section useful, and hence based the analysis steps around that model.

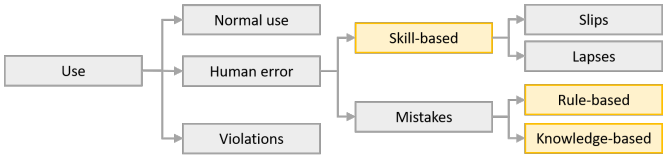

Fig. 3. Human Error according to [19] (grey terms) and [18] (yellow terms).

For automated driving, we use the terminology of SAE J3016 [6]. Some of the key terms such as ADS and DDT have already been introduced. Another key concept is ODD, which

is defined as the collection of operating conditions where a specific ADS feature is designed to be used. Other terms we use are human driver (HD) and human user (HU). These and other key terms and their acronyms are summarized in Table I; some of the terms are further introduced later in the paper.

TABLE I
LIST OF ACRONYMS

Terms from SAE J3016	
ADS	*Automated Driving System* - automation on SAE levels 3-5.
DDT	*Dynamic Driving Task* - the operational and tactical functions needed to operate a vehicle in traffic.
HD	*Human Driver* - user who performs part or all of the DDT and/or DDT fallback.
HU	*Human User* - human in the role of either driver, passenger, DDT fallback-ready user or dispatcher for driverless operation.
ODD	*Operational Design Domain* - operating conditions under which an ADS feature is designed to function.
Terms used in functional safety / dependability	
CCA	*Cause-Consequence Analysis*.
FTA	*Fault Tree Analysis*.
FMEA	*Failure Modes and Effects Analysis*.
HARA	*Hazard Analysis and Risk Assessment*.
SOTIF	*Safety of the Intended Functionality*.
E/E	*Electrical/Electronic*.
Terms used in human factors	
SA	*Situation Awareness* - perception and comprehension of the environment, and projection of its future status.
HMI	*Human-Machine Interaction* (or *Interface*).

B. Human Behaviour Analysis

The situation awareness (SA) model of Endsley [17] is generally considered the most used SA model [20], and have unsurprisingly been under scrutiny, which has been reviewed by the author [21]. The cyclical model allows us to analyze the situation awareness and dynamic decision making of the HU, illustrated in Fig. 4. The SA model describes how an individual's perception (P), comprehension (C) and projection (PR) of future state of a situation forms the ascending three level basis of their SA. The SA is followed by action selection, or decision (D), and performance of action (A). Individual and task/system factors affect each step in the decision making process.

Just as there are several variations of what defines an error, there are several ways of classifying human errors, which is why the most practical utility for the analysis should decide [19]. Focus could be at task/system factors (e.g. priming, stresses or interruptions), or at the outcome on the state of the environment after performed actions (e.g. incidents or accidents). The safety analysis in this paper focuses on the intent and actions of the human in each stage of the decision making process. For the purpose of providing a comprehensive safety analysis of the human behaviour of the human-ADS interaction, each stage in the SA model should be analyzable for potential errors.

Jones and Endsley [22] uses the SA model to study which levels that are the most common source of SA related errors in aviation. The study is limited to the three levels of SA, defining

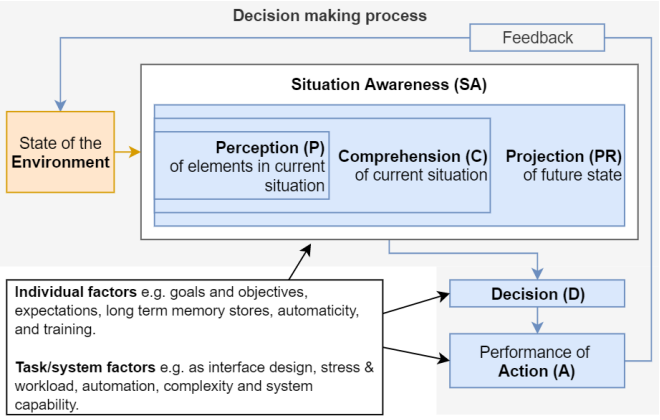

Fig. 4. A compressed visualisation of the model of situation awareness in dynamic decision making [17] used as part of the safety analysis.

the deviation of each part the SA consists of, not going into detail of errors for Decision and Action. However, Hollnagel [12] suggests a classification that complements Jones and Endsley [22] with two definitions of deviations of decision and action. Together the two studies cover each stage in the decision making process, as seen below:

- Perception (P) - failure to correctly perceive the information [22].
- Comprehension (C) - failure to comprehend the situation [22].
- Projection (PR) - failure to project the situation into the future [22].
- Decision (D) - Incorrect selection of action to reach a goal, or incorrect execution of that action [12].
- Action (A) - Unintentional substitution of a correct performance segment (action) with an incorrect one [12].

The human behavior analysis based on SA would be aided by a compilation of SA errors for each of the stages P/C/PR/D/A relevant for the automotive domain to guide the analyst. While this is an area that requires further research, work such as the comprehensive overview by Stanton and Salmon [23] could be used as one of the sources for such a guide, even if it is not organized according to the SA model.

C. Relation to Automotive Standards

In this section we discuss how the interaction framework in general and the process proposed in this paper in particular can relate to the most established safety standard in the automotive domain, ISO 26262:2018 - Road Vehicles - Functional Safety [2], which is applicable for series production road vehicles such as cars, trucks, buses and motorcycles. The standard itself defines functional safety as *absence of unreasonable risk due to hazards caused by malfunctioning behavior of E/E systems*. In addition, we similarly discuss the newer ISO 21448:2019 PAS - Safety of the Intended Functionality (SOTIF), which is intended as a complement to ISO 26262.

1) ISO 26262:2018 - Functional Safety - Road Vehicles: ISO 26262 provides a safety lifecycle and a risk-based ap-

proach to determine necessary risk reduction in terms of discrete levels called Automotive Safety Integrity Levels (ASIL), and which requires risk reduction strategies in terms of safety mechanisms and stringency in development. The hazard analysis and risk assessment (HARA) involves identifying potential hazardous events, which are then are assigned an ASIL from A to D based on a risk assessment. ASIL D indicates the most critical hazardous events, and A the lowest. The exposure, potential severity, and driver controllability of the identified events are considered when determining the necessary ASIL. A hazardous event could also be considered not safety-critical at all, so that no additional safety mechanisms are necessary (e.g. if exposure is considered to be 0). Top-level safety requirements, called safety goals, are defined to cover all hazardous events with an assigned ASIL and inherits the ASIL of the corresponding hazardous event(s).

The achievement of functional safety is shown in a *safety case*, which is an argument, supported by evidence, that functional safety is achieved. This includes showing that the safety goals are correctly elicited and implemented. A note in the standard clarifies that the *"safety case can be extended to cover safety issues beyond the scope of ISO 26262"*. This would mean that the interaction framework could be included as it is a vital part of safety for e.g. many ADS features. From an ISO 26262 perspective, the interaction framework and the HU could potentially also be treated like an *external measure*, which is a measure external to the item (i.e. function to which ISO 26262 is applied) which reduces or mitigates risks in the item. It is of course imperative that such measures do not adversely affect safety, i.e. one needs to make sure the external measure is at least as good as indicated by the safety goals it is affected by. The potential risk mitigation of external measures, e.g. increasing the controllability of the HU as discussed in [24], can be accounted for in the hazard analysis. Functional safety requirements, which are derived from the safety goals, shall also be derived for external measures and ISO 26262 is still applicable for measures implemented as E/E functions. For the interaction framework, only the HMI will typically be part of the E/E system. We need other analysis techniques to make sure the human contribution does not adversely affect safety. There may be a dependency between HU and E/E systems that is not considered in the common analysis methods used in the functional safety community. This is illustrated in Fig. 5, e.g. the failure by the HU to correctly perceive a signal from the ADS may be due to bad design of a HMI component that is part of the E/E system.

2) ISO 21448:2019 PAS - SOTIF: While ISO 26262 provides a comprehensive safety lifecycle, its scope means guidance on some safety-critical issues are missing, one of which are potential safety problems concerning the human actions (or inactions) when interacting with the vehicle HMI. A complementary specification, ISO 21448:2019 PAS - Safety of the intended functionality (SOTIF), has recently been released[1].

[1]At the time of writing, work is ongoing to develop this publicly available specification (PAS) to a full standard. This paper considers only the content of the currently existing 21448:2019 PAS.

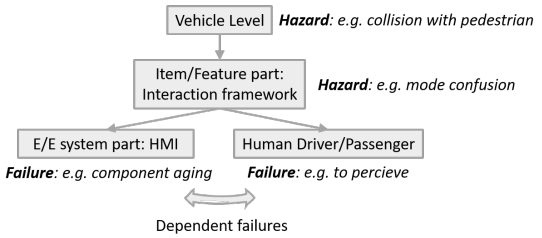

Fig. 5. Hazards on vehicle level and interaction framework. There is a dependency between HMI and human users.

ISO 21448 PAS has a slightly different approach, focusing on identifying weaknesses in the functional specification and resolving safety issues by functional modifications. The focus is primarily performance limitations in technologies such as environment sensors providing situational awareness for the E/E system (including machine learning components), but also HMI related hazards. ISO 21448 PAS denotes all use of a system in a way not intended by the manufacturer *misuse* and scenarios describing such potential actions *misuse scenarios*. Fig. 6 illustrates a method for deriving such misuse scenarios described in the informative annex E of ISO 21448 PAS. Contrary to what the term might indicate, misuse does not include deliberate violations, only human errors.

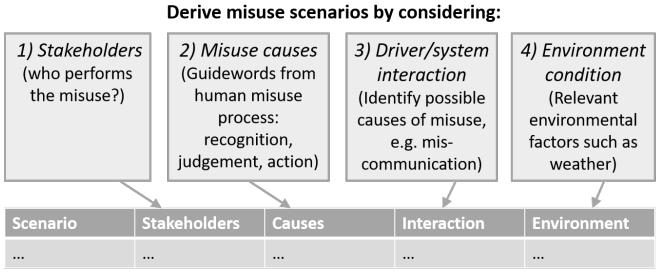

Fig. 6. ISO 21448 PAS method (informative) for identifying misuse (human error) scenarios.

Rather than a safety lifecycle, ISO 21448 PAS is built around a workflow containing the following main activities:

- Hazard analysis and risk evaluation - focusing on the scope discussed above.
- Identification of triggering events for these hazards.
- Functional modification to reduce identified risks.
- Definition of verification and validation strategy to determine that SOTIF hazards are adequately addressed.
- Criteria for release including review of SOTIF activities and evaluation of the acceptability of residual risk.

While the ISO 21448 PAS includes requirements for treating HMI related hazards, and provides an example process for identifying misuse scenarios (Fig. 6), it does not provide a methodology for structured analysis and risk assessment of a HMI. The interaction safety analysis process described in this paper could be one way to fulfill the requirements regarding human interaction in ISO 21448 PAS, and provide part of the safety argumentation for the HMI needed as criteria for

release. Missing in our method, however, is verification and validation strategy. ISO PAS 21448 refers to the Human Factors Analysis and Classification System analysis [25], which is based on Reason [9].

D. Automotive vs. Other Domains

As a short note on the relation to other domains, much work on human interaction has undoubtedly been done in other safety-critical domains such as avionics and nuclear, for instance work on automation surprises [26] or methods for assessing the interrelation between dependability and usability [27] just to name a few. When putting the concepts discussed in this paper in perspective compared to this work there are important differences in fundamentals, demanding new approaches. This includes how well educated and familiar the users are with the systems, the responsibility agreement and the time-frames within the different contexts (time available for corrective action is typically much shorter in the automotive domain). Using avionics as an example the user is always an educated pilot and there are procedures in place to mitigate risks. For passenger cars, while typically requiring a drivers license, relying on the driver to learn and perform lengthy procedures for risk mitigation is more precarious. While beyond the scope of the paper to explore, one can also note that passenger cars are also, in contrast, consumer products, which implies a vastly different legal landscape.

III. Case Study

We will demonstrate the proposed methodology using a running example of a protocol used for transferring the control from an HD to an ADS. Fig. 7 shows different states of an ADS equipped vehicle. We assume this ADS feature can be enabled and disabled while the vehicle is in motion in normal traffic situations, that is, the HU may be performing the DDT for parts of the trip (HD-DDT) while the ADS performs the DDT for other parts (ADS-DDT). The ADS have further states for responding to conditions where it can no longer fulfill its original strategic task, the fallback state whose purpose is to reach a minimal risk condition. All the ADS-DDT states need to be within the ODD.

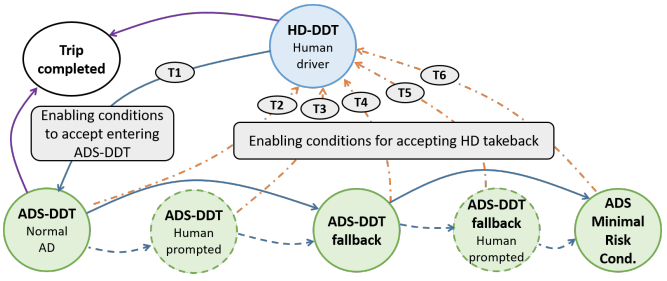

Fig. 7. ADS states and transitions for an example level 4 system conforming to SAE J3016 [6] (see figures 5-7 in J3016:JUN2018).

This model results in a number of transitions between the HD and ADS each with slightly different conditions and challenges. The transitions are labelled *T1* to *T6* in 7, where *T1* is a transition from the HD to the ADS and the others are transitions from the ADS to a HU. Along the state transition edge for *T1* is *Enabling conditions to accept entering ADS-DDT*, which are the conditions within the ODD where a transition from HD to ADS is possible. Note that these can be (in fact probably are) more restrictive conditions than what are possible for the ADS normal operation. E.g. a transition may not be allowed in the middle of an overtake action even if both HD and ADS by themselves could handle the overtake, this in order to avoid *unfair transition* hazards. Thus, for a transition to actually occur, there needs to be a protocol, where the HD and ADS communicate in such a way as to allow for a transition to complete while avoiding transition hazards.

The proposed protocol for *T1* in our example, which is slightly modified from a previous paper [7] from the same project[2], is shown in Fig. 8. Starting from HD-DDT, state *A*, the ADS makes a transition to *B* when enabling conditions for entering ADS-DDT are fulfilled, i.e. when it is possible for the HD to initiate a transition. The transition then requires two actions from HD with an acknowledge response from the ADS in between. If any step of the protocol is not completed while the enabling conditions are fulfilled, the protocol reverts to *A*. In this paper we focus on this transition only, but note that the ADS to human transitions *T2-T6* have additional difficult failure modes pertaining to the readiness of the HU to take control. Another important property of the example is that control is always given, never taken. That is, neither HU nor ADS can forcibly take control of the vehicle, transfer of control is only possibly by mutual agreement by use of the transition protocol.

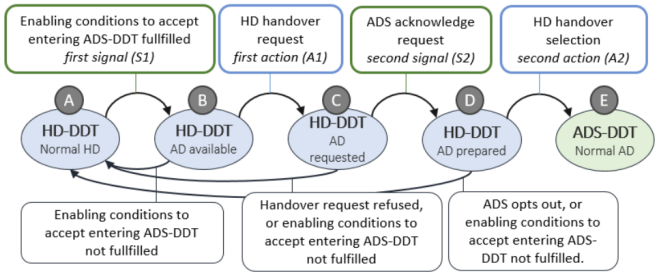

Fig. 8. Proposed protocol for transition T1 (HD handover to ADS).

An example of how the transition protocol *T1* in Fig. 8 could be designed:

- State change *A to B*: Tell-tale light "ADS available"
- State change *B to C*: Push of button as first HD action
- State change *C to D*: Tell-tale light "ADS prepared"
- State change *D to E*: Change of lever as second HD action (lever is locked until ADS is in prepared state)

It is evident that this proposed protocol already have several features meant to reduce the risk of transition hazards. For instance, the requirement for two actions to complete the transition is meant to guard against the HD enabling or disabling the ADS by accidentally pushing the wrong button,

[2] The ESPLANADE project, see https://esplanade-project.se.

and the property that control is always negotiated in mutual agreement is to ensure there are no unfair transitions. The analysis of interactions should reveal if the proposed protocol is sufficiently safe or if there are remaining failure modes that it does not offer acceptable protection against.

IV. Analysis of Interactions

This section will go through and explain the proposed process, illustrated in Fig. 9, by using the case study to exemplify each step. For the case study, we assume the available HMI specification (step 1) is only what was stated in Sec. III. It should be noted that this is a somewhat simplified specification, in reality some more information on e.g. enabling conditions, timeouts and design of the HMI elements would increase the precision of the analysis.

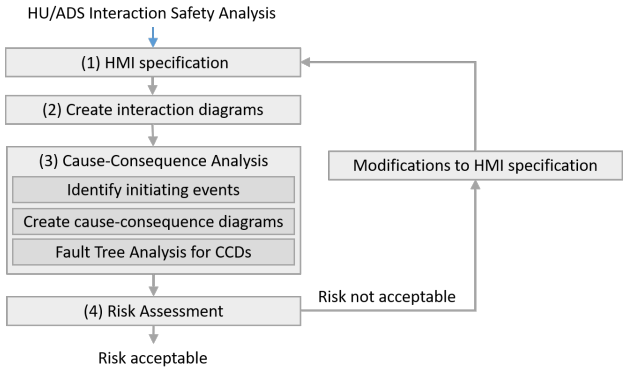

Fig. 9. Overview of the proposed interaction safety analysis process.

A. Interaction Sequences

We visualize both intended use and deviations of the transition protocol in a manner inspired by sequence diagrams, with the ADS and HU as two communicating agents with the HMI between. The intended interaction sequence of *T1* is shown in Fig. 10. Each step in the decision making process from the SA model of Sec. II-B is included to allow for later fault analysis. Arrows indicate signals and actions that go via the HMI - green arrows and boxes for actions initiated by the ADS and blue arrows and boxes for the HU. Each agent's comprehension of the current protocol state (from Fig. 8) is also indicated, and which agent that is actually performing the DDT.

To assist the reader in how the interaction sequence visualization works Fig. 11 shows how the same protocol, i.e. *T1* from the case study, works in terms of the HMI elements (tell-tales, button and lever) described in Sec. III.

Additional interaction sequence diagrams should be used to model transition protocol deviations based on the protocol sequence. For instance, Fig. 12 shows a sequence when both signals from the ADS (ADS available and ADS prepared) are missing. The first point where one of the agents confuses the current protocol state is highlighted with yellow and labelled *transition protocol confusion*. This deviation could, if not identified and mitigated in the protocol, lead to the hazard *mode confusion*, highlighted in red. The dotted green arrows

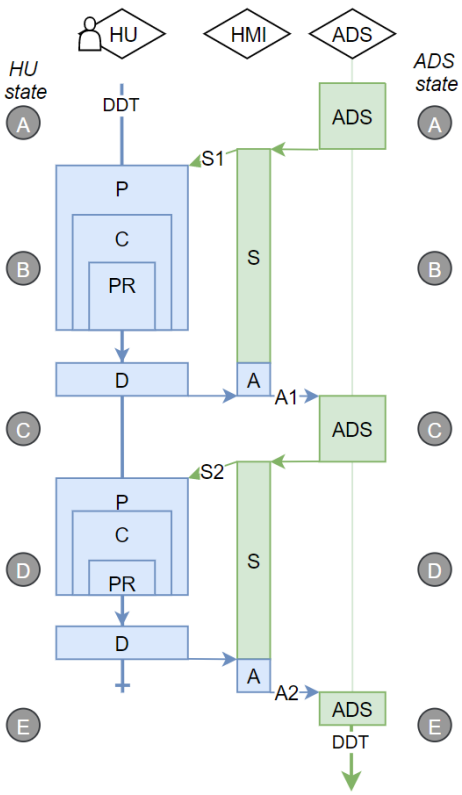

Fig. 10. The intended interaction sequence for T1.

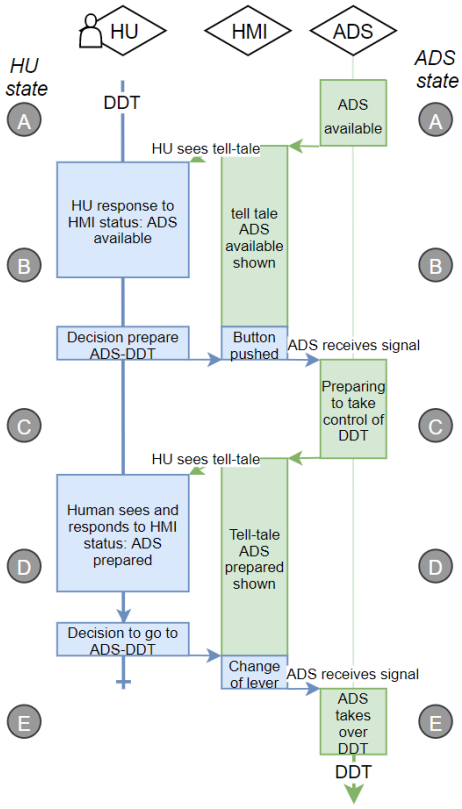

Fig. 11. Descriptive example of the intended interaction sequence for *T1*.

and box outlined from the HMI to HU indicates analyzable failures, classified in Sec. II-B. For example, it could be that the HU perceives the tell-tale for ADS available, even if it is actually not on. The failed perception leads to a comprehension and projection that is not aligned with the ADS real state (that ADS is not available). The HU could then decide to enter ADS available, and correctly perform the action to reach that goal, but due to the failed perception it is not possible, and would (if not somehow mitigated) lead to *driving mode confusion*. The root causes can be analyzed in the corresponding cause-consequence diagram and fault tree analysis, which will be discussed in the next section.

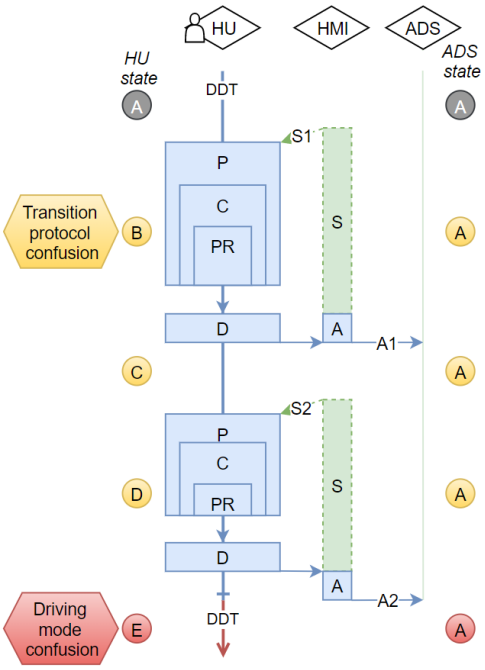

Fig. 12. Interaction sequence for a mode confusion scenario.

The issue of completeness is always difficult. Did we find all relevant deviations from the intended sequence? A tool to aid the analysis is to consider the possible ways intervals can overlap [28], as illustrated in Fig. 13. Also note that the magnitude of e.g. interval overlap or gap can be relevant. For instance, for the signaling between the two agents there will always be some time gap as it takes time for e.g. the HU to perceive a signal from the ADS. Here the relevant question is whether there is a specific tolerance margin which, if surpassed, could lead to a failure. In our protocol example we have only considered a well-defined transition point where the authority over the DDT switches from one agent to the other (*x meets y* in Fig. 13). Protocols where the control of the DDT is shared during some period of time are possible, although since both safety analysis and the responsibility/liability issue if an accident occurs will be tricky we have currently not considered these. Similarly to how all safety analyses are treated in ISO 26262, we also recommend review by independent persons to reduce the likelihood of omissions in the analysis.

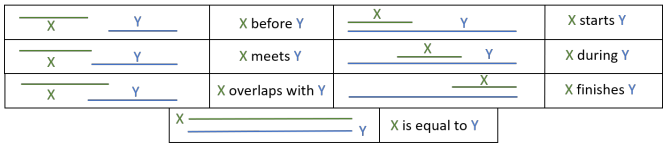

Fig. 13. Possible combinations of two intervals.

B. Cause-Consequence Analysis

There are several well known analysis methods for identifying root causes for failures, the most known probably being failure modes and effects analysis (FMEA) and fault tree analysis (FTA). However, transition protocols can include a sequence of actions and signals through the HMI. Therefore we have instead opted to use cause-consequence analysis (CCA) [29]. It was developed to analyse potential accident scenarios consisting of a sequence of events starting with an initiating event (IE) and proceeding through a number of intermediate events to an eventual outcome. Based on each identified initiating event and its associated intermediate events, a cause-consequence diagram (CCD) is created; the diagram will form a tree from the initiating event, branching for each intermediate event based on whether the outcome is as expected or not. For each unsuccessful outcome, an FTA can be made to find root causes for that unsuccessful outcome. In our case creating the fault trees will include the human behavior analysis presented in Sec. II-B in order to include faults for the human interaction.

1) Identify Initiating Events: Identification of accident scenarios and from these initiating events should be done with hazard analysis or system assessment [29]. In our method this system assessment is based on the interaction sequences, which have already identified the accident scenarios and their possible initiating events. As we want to analyze the interaction framework rather than the entire function, accident scenarios will be scenarios leading to one of the interaction hazards. Thus the initiating events will be at the points when transition protocol confusion occurs, i.e. the first event where the belief of protocol state of HU and ADS differs. As intermediate events, we use the well-defined observable events in the protocol, that is stimuli from the HMI to HU (S1 and S2) and actions from HMI to ADS (A1 and A2).

Table II lists initiating events elicited from an analysis of the *T1* protocol of Fig. 8; it considers the interaction sequence in Fig. 12 as well as three other mode confusion scenarios where the transition protocol confusion occurs later in the protocol. Note that this example does not provide an exhaustive analysis; there might be other relevant IEs based on interaction sequences with combinations of intervals not included in the example, e.g. stimuli or actions happening to early or too late.

2) Create Cause-Consequence Diagrams: A separate CCD is created for each identified IE in the proposed protocol. Fig. 14 shows CCDs for the IEs listed in Table II. CCDs are relatively straight-forward. As mentioned, a tree is formed

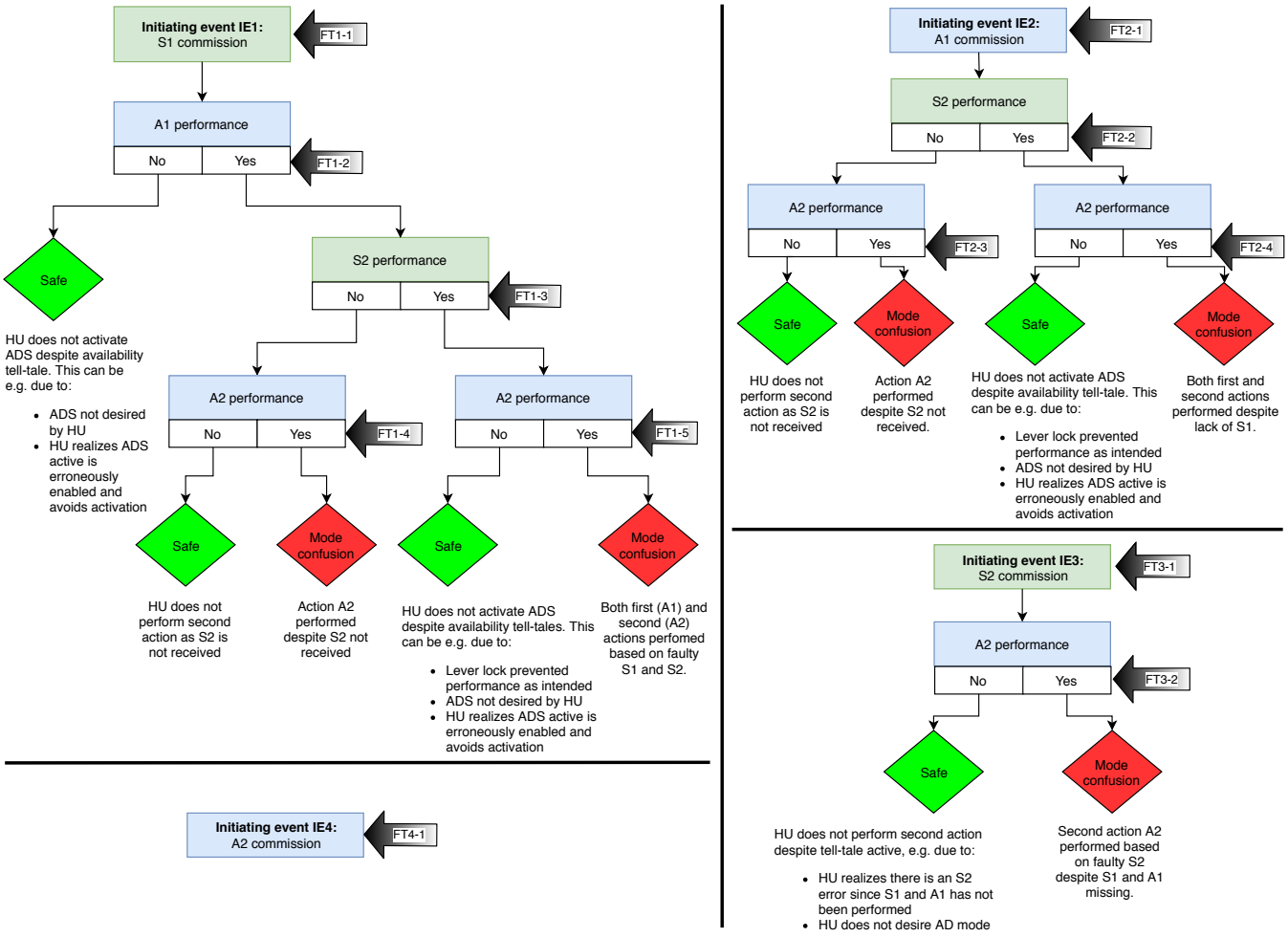

Fig. 14. Cause-consequence diagrams for the initiating events of the mode confusion scenario.

TABLE II
EXAMPLE ANALYSIS OF INITIATING EVENTS FOR PROTOCOL T1.

IE#	Initiating Event	Explanation
IE1	S1 commission	S1 incorrectly provided.
IE2	A1 commission	A1 performed without correct S1.
IE3	S2 commission	S2 incorrect provided.
IE4	A2 commission	A2 performed without correct S2.
IE5	.....	.....

with the initiating event as the root, and each initiating event fans out depending on the outcome of the intermediate event. It is important, however, to keep in mind what the desired outcome is when constructing the CCDs and associated FTs. For instance, in the CCD for IE1 in Fig. 14, the initiating event is a commission (undesired activation) of stimuli 1 (S1). Since S1 is faulty, the desired outcome here is that A1 (first action to activate the ADS - push of button) does not occur. Hence, the fault tree for the intermediate event ”A1 performance” will be for the outcome ”Yes”, as this is the outcome which can potentially lead to a hazard as the final outcome. We formulate the intermediate events neutrally since either ”Yes” or ”No” may lead to an undesired outcome depending on the sequence of events. Likewise, it is important to remember the initial state. For the IE2 CCD, the initiating event is A1 commission. This means we assume everything prior to the IE has transpired as expected. As we have an A1 commission, in this case it means A1 has been activated even though S1 had not occurred.

3) Fault Tree Analysis for Undesired Outcomes of Events: Fault tree analysis for three of the nodes potentially leading to hazards are shown in Fig. 15. As can be seen, we include both E/E and human faults in the same FT, as the same top event can sometimes be caused by either. In this case we have concluded in FT1-1 that S1 commission is a faulty tell-tale with a couple of potential causes. For the example we have not developed the fault tree beyond this level[3]. It can also be noted that, as we have chosen S1 commission as initiating event, we do not include errors due to human SA in this fault tree. These types

[3]A diamond-shaped symbol beneath a node means the fault could be developed (broken down) further. A circle means it is a basic event, i.e. failure in a component with no other underlying cause. For space reasons we draw these symbols beneath the textbox rather than having the text within the diamond/circle, which is also a common way to draw FTs.

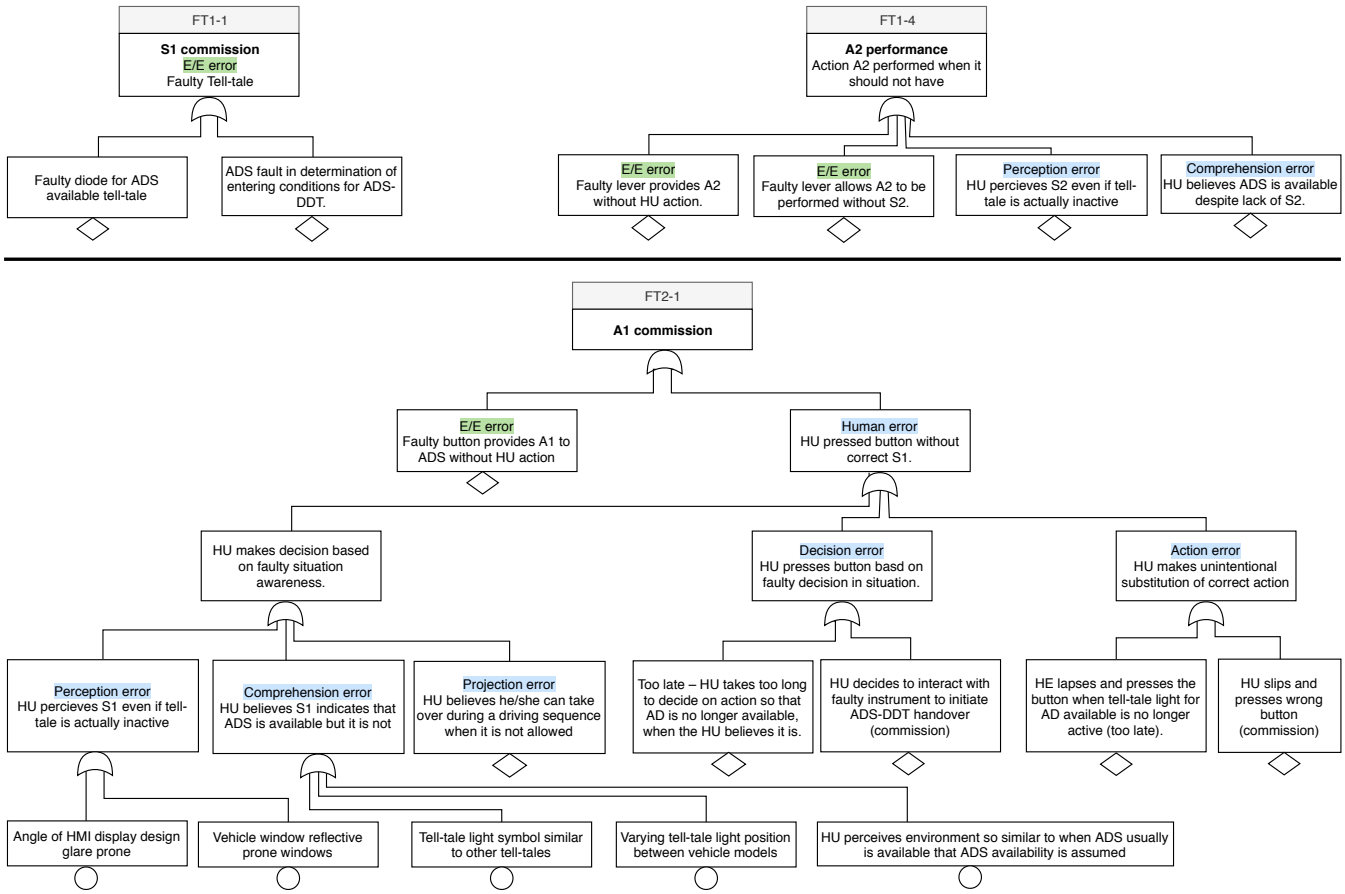

Fig. 15. Some of the fault trees from analysis of CCD nodes leading to hazards for the mode confusion scenario.

of errors are instead included in FT2-1, where the initiating event is A1 commission, i.e. the push-button was activated without the tell-tale being active. Thus, in FT2-1, human errors such as erroneous perception or comprehension of the tell-tale are included as potential causes for the A1 commission fault. FT 1-4 shows another example where either E/E or human error could cause the undesired outcome.

C. Risk Assessment and Risk Reduction

CCA is originally intended to be combined with probabilistic risk assessment, i.e. where a probability can be assigned to each event, as it was developed for systems where these probabilities can be obtained. However, this is typically not the case for our analysis. So the question is how should we use the analysis for risk reduction? If the interaction framework is treated as an external measure in ISO 26262, the E/E parts would get assigned safety requirements with ASILs stemming from the vehicle level hazard analysis (refer back to Fig. 5). However, the human errors are not solved by the ISO 26262 lifecycle. Instead, we should make sure to design the HMI to sufficiently reduce the risk from human error. But this begs the question, what is sufficient? How do we relate it to the risk reduction requirements of e.g. ISO 26262?

This is an area where we believe further research is needed. One option would be to use a qualitative assessment similar to criticality ranking used in FMECA [3] where risk is assessed as a function of severity and likelihood with a ranking that might look like Fig. 16. Due to the difficulty assessing the parameters we kept the matrix very simple, and as probabilistic analysis in not possible, this assessment must be applied individually for all human faults in the FTs for each event. Based on the ranking, all risks considered critical would require risk reduction, e.g. a redesign of the HMI to make it more resilient against these faults.

		Severity	
		Low	*High*
Likelihood	*Improbable*	Non-critical	Non-critical
	Low	Non-critical	Critical
	High	Critical	Critical

Fig. 16. Criticality matrix.

For instance, in FT2-1, the likelihood of perception error might be low, but severity given a possible outcome of mode confusion (which if we assume it could in turn lead to e.g. a collision with a cyclist based on a vehicle level HARA of the feature) is high. Therefore, this is considered critical and requires some risk reduction, e.g. redesign of the HMI using

multi-modal information instead of just a light to bring the likelihood of perception errors to improbable.

V. Conclusions and Future Work

In order to make a complete safety case for an automated driving system, a safety analysis of the interaction between humans and the ADS is necessary. However, there is currently no consensus on how to make such an analysis in a way so that both functional safety and human factors expertise can contribute. This paper proposes a method that may provide a step on the way, and even if the usage of the interaction safety analysis method would likely be applicable in other domains than automotive, e.g. machinery, we attempt to place the method in context of two automotive safety standards, ISO 26262:2018 and ISO 21448:2019 PAS.

The well established taxonomy of [8] from the dependable systems community has been compared with well-known work within the human factors community [9], [12], [18] in regards to how "human error" is viewed, and differences discussed. This paper shows the need for, and takes a step towards synchronization of the two communities' terminology. Through a common terminology faults can be better identified, discussed and improvements made towards creating a safe HMI and interaction sequence for safe transitions.

Endsleys model [17] is not necessarily the only plausible option for modelling human behaviour, but the ascending levels of situation awareness, and cyclic decision making process was very practical, both to synchronize with the sequence diagrams and to break down, and step by step analyze, the possible human errors through FTA. However, by analyzing the initiating events for the proposed protocol (listed in Sec. II) through the cause-consequence diagram, application of different human behaviour models is possible. We encourage further research to test not only the proposed methodology, which is not done in this study, but also using other models for human behaviour.

Some questions we aim to answer in future work, e.g. how to accurately capture risk of successive mistakes or stuck in transition hazards resulting from too long or complicated sequences. As the CCD is best at capturing faults contained at each event this is not obvious. We would also like to explore how general the method is with regards to other human-ADS interactions beyond transition protocols, and how it can be used together with other techniques such as [4].

Acknowledgment

The authors would like to thank Kenneth Östberg for valuable input and discussions.

References

[1] P. Koopman and M. Wagner, "Challenges in Autonomous Vehicle Testing and Validation," *SAE International Journal of Transportation Safety*, vol. 4, no. 1, pp. 15–24, Apr. 2016.

[2] ISO, "ISO 26262:2018 Road vehicles – Functional safety," 2018.

[3] U.S. Department of Defense, "MIL–HDBK–882D: Standard Practice for System Safety," 1998.

[4] C. Martinie, P. Palanque, R. Fahssi, J.-P. Blanquart, C. Fayollas, and C. Seguin, "Task model-based systematic analysis of both system failures and human errors," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 2, pp. 243–254, 2015.

[5] M. Skoglund, F. Warg, and B. Sangchoolie, "Agreements of an automated driving system," in *37th International Conference on Computer Safety, Reliability, & Security (SAFECOMP 2018) - Fast Abstract*, Vasteras, Sweden, Sep. 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01878603

[6] SAE, "SAE J3016 - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," 2018.

[7] R. Johansson, J. Nilsson, and A. Larsson, "Safe transitions between a driver and an automated driving system," *International Journal on Advances in Systems and Measurements*, vol. 10, no. 3-4, 2017.

[8] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, "Basic concepts and taxonomy of dependable and secure computing," *IEEE transactions on dependable and secure computing*, vol. 1, no. 1, pp. 11–33, 2004.

[9] J. Reason, *Human Error*. Cambridge University Press, 1990.

[10] ISO, "ISO 62366: 2008 medical devices–application of usability engineering to medical devices," 2008.

[11] G. Salvendy, *Handbook of human factors and ergonomics*. John Wiley & Sons, 2012.

[12] E. Hollnagel, "Human error," in *Position paper for NATO conference on human error*, 1983.

[13] M. S. Donaldson, J. M. Corrigan, L. T. Kohn *et al.*, *To err is human: building a safer health system*. National Academies Press, 2000, vol. 6.

[14] E. Hollnagel and D. D. Woods, *Joint cognitive systems: Foundations of cognitive systems engineering*. CRC Press, 2005.

[15] E. Hollnagel, O. Pedersen, and J. Rasmussen, "Notes on human performance analysis," 1981.

[16] S. Dekker, *The field guide to understanding'human error'*. CRC press, 2017.

[17] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.

[18] J. Rasmussen, "Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models," *IEEE transactions on systems, man, and cybernetics*, no. 3, pp. 257–266, 1983.

[19] J. Reason, *The Human Contribution: Unsafe Acts, Accidents and Heroic Recoveries*. Routledge, 2008.

[20] D. Golightly, J. R. Wilson, E. Lowe, and S. Sharples, "The role of situation awareness for understanding signalling and control in rail operations," *Theoretical Issues in Ergonomics Science*, vol. 11, no. 1-2, pp. 84–98, 2010.

[21] M. R. Endsley, "Situation awareness misconceptions and misunderstandings," *Journal of Cognitive Engineering and Decision Making*, vol. 9, no. 1, pp. 4–32, 2015.

[22] D. G. Jones and M. R. Endsley, "Sources of situation awareness errors in aviation." *Aviation, space, and environmental medicine*, 1996.

[23] N. A. Stanton and P. M. Salmon, "Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems," *Safety Science*, vol. 47, no. 2, pp. 227–237, 2009.

[24] M. Sassman and R. Wiik, "Safer transitions of responsibility for highly automated driving: Designing HMI for transitions with functional safety in mind," in *10th European Congress on Embedded Real Time Software and Systems (ERTS 2020)*, Toulouse, France, Jan. 2020.

[25] S. A. Shappell and D. A. Wiegmann, "The human factors analysis and classification system–HFACS," 2000.

[26] N. B. Sarter, D. D. Woods, C. E. Billings *et al.*, "Automation surprises," *Handbook of human factors and ergonomics*, vol. 2, pp. 1926–1943, 1997.

[27] C. Fayollas, C. Martinie, P. Palanque, Y. Deleris, J.-C. Fabre, and D. Navarre, "An approach for assessing the impact of dependability on usability: application to interactive cockpits," in *2014 Tenth European Dependable Computing Conference*. IEEE, 2014, pp. 198–209.

[28] J. F. Allen, "Maintaining knowledge about temporal intervals," in *Readings in qualitative reasoning about physical systems*. Elsevier, 1990, pp. 361–372.

[29] C. A. Ericson, *Hazard analysis techniques for system safety*. John Wiley & Sons, 2015.