Requirements for Monitoring Inattention of the Responsible Human in an Autonomous Vehicle: The Recall and Precision Tradeoff

Johnathan DiMatteo Daniel M. Berry
School of Comp. Science
University of Waterloo
Waterloo, ON N2L 3G1, Canada
jdimatte@uwaterloo.ca dberry@uwaterloo.ca

Krzysztof Czarnecki
Dept. of Elect. & Comp. Engg.
University of Waterloo
Waterloo, ON N2L 3G1, Canada
k2czarne@uwaterloo.ca

Abstract

Recent fatal accidents with partially autonomous vehicles (AVs) show that the responsible human in a vehicle (RHV) can become inattentive enough not to be able to take over driving the vehicle when it gets into a situation that its driving automation system is not able to handle. Studies show that as the level of automation of an AV increases, the tendency for the RHV to become inattentive grows. To counteract this tendency, an AV needs to monitor its RHV for inattention and when inattention is detected, to somehow notify the RHV to pay attention. Requirements engineering for the monitoring software needs to trade off false positives (FPs) and false negatives (FNs) (or recall and precision) in detecting inattention. An FN (low recall) is bad because it represents not detecting an inattentive RHV. An FP (low precision) is bad because it leads to notifying the RHV too frequently, to the RHV's ignoring notifications, and thus to degraded effectiveness of notification. The literature shows that most researchers just assume that FPs and FNs (recall and precision) are equally bad and weight them the same in any tradeoff. However, if, as for aircraft pilots, notification techniques can be found whose effectiveness do not degrade even with frequent repetition, then many FPs (low precision) can be tolerated in an effort to reduce the FNs (increase the recall) in detecting inattention, and thus, to improve safety.

1 Introduction

Safety-critical systems in nuclear power, medicine, and transportation rely on vigilant operators to guarantee low risk. However, as automation replaces traditional human roles, operators are forced to adapt and develop new skills. Whether or not this transformation will lead to greater system safety is not yet clearly established. A National Transportation Safety Board (NTSB) report found that in 31 of the 37 serious accidents involving U.S. air carriers from 1978–1990, inadequate attention played a major role [6]. Pilots or crew members neglected to govern instrumentation, verify inputs, and communicate caught errors. During the period of review by the NTSB, the aviation industry was undergoing significant

In: M. Sabetzadeh, A. Vogelsang, S. Abualhaija, M. Borg, F. Dalpiaz, M. Daneva, N. Fernández, X. Franch, D. Fucci, V. Gervasi, E. Groen, R. Guizzardi, A. Herrmann, J. Horkoff, L. Mich, A. Perini, A. Susi (eds.): Joint Proceedings of REFSQ-2020 Workshops, Doctoral Symposium, Live Studies Track, and Poster Track, Pisa, Italy, 24-03-2020, published at http://ceur-ws.org

changes in automation levels. As the responsibility for flying shifted from the pilot to the software, pilots were lulled into a false sense of confidence. Pilots felt bored and abdicated their duties.

Today, the automotive industry is experiencing a similar phenomenon, highlighted by two recent fatalities involving cars with different degrees of autonomy.

(1) On a dark night in March 2018, an Uber Technologies vehicle, being tested operating fully autonomously, struck and killed a street-crossing pedestrian. The backup driver, responsible for commandeering and then operating the vehicle in times of emergency, was distracted moments before the incident [11]. Her hands were not on the steering wheel, her eyes were directed downward just before impact, and she did not engage the emergency brakes. The report concludes that the vehicle's operating software had apparently recognized the pedestrian as something else.

(2) In March 2019, only ten seconds after a driver initiated the only partially autonomous driving mode, ironically called "Autopilot", in a commercially available Tesla Model 3, the Tesla struck the underside of a semi-trailer, killing the driver instantly. No evasive maneuvers had been executed. The driver was apparently not paying attention [12]. His hands had not been on the steering wheel, as is required by the use of Tesla's Autopilot.

Each incident demonstrates the willingness or the tendency for a human to withdraw attention while he or she is supervising or driving a more or less autonomous vehicle. The inattention problem, the same that had plagued the aviation industry many years earlier, had returned.

Over the next few decades, the automotive industry will experience several challenges caused by the increasing autonomy in vehicles. This increasing autonomy is moving more and more of the vehicle's operation into the vehicle's software, causing the requirements of the vehicle to change and shifting the driving responsibility from the driver to the software, seemingly allowing and thus, encouraging the driver to focus his or her attention elsewhere. However, no vehicular software is perfect, and until such time as it is close enough to being perfect, some human being, must be responsible for continually paying attention to the AV and its surroundings to be able to take over for an AV that is about to get into trouble. Therefore, many are considering equipping each AV with machine-learning (ML) based artificial intelligence (AI) that monitors the attentiveness of the *responsible human in a vehicle (RHV)* that is supposed to be driving or supervising[1] the AV. An RHV is *supervising* a vehicle when the RHV is *watching over the vehicle enough to be able to take over driving at any time, sometimes in an emergency, sometimes at the vehicle's request*.

In general, a *vehicle*, which can be a car, van, SUV, or truck, is *autonomous* to *some degree*, and it requires human *attention* to *some degree*. The Society of Automotive Engineers (SAE) has defined in its J3016 standard, six autonomy levels with which to classify the degree of autonomy of a vehicle according to (1) the degree of human attention required and (2) the degree of automation of the vehicle [14]. This paper focuses on two particular levels of autonomy:

Level 2 in which there is an RHV tasked with driving the vehicle even though some of the functions of the vehicle are automated; that is, the RHV must keep his or her hands on the vehicle's steering wheel at all times and is responsible for operating the vehicle at all times *even when the vehicle is doing a function that has been automated*; and

Level 3 in which there is an RHV tasked with supervising the vehicle even as the partially autonomous vehicle is doing many functions; that is, while the RHV does not have to keep his or her hands on the vehicle's steering wheel at all times, he or she must be attentive enough about the vehicle's current condition to be able to take over driving the vehicle at *any* time, sometimes in an emergency, and sometimes at the vehicle's request.

Thus, the RHV is (1) a driver or (2) a supervisor, and what an RHV does, *piloting*, is (1) driving or (2) supervising.

1.1 RHV Monitor and Notifier Because an RHV's *attention* may lapse while he or she is piloting an AV, among the vehicle's software must be a *RHV Monitor and Notifier (RMN)*. The RMN consists of two communicating parts:

(1) the *Monitor*, an AI that somehow *monitors* the RHV for signs of *inattention*, and at any time that the Monitor *detects* that the RHV is inattentive[2], it *informs* the Notifier to do its job, and

(2) the *Notifier*, when informed by the Monitor, somehow *notifies* the AV[3], the RHV, or both, that signs of inattention have

[1] Unfortunately, the literature uses the word "monitoring" to describe both (1) the act of watching over an AV to detect when the AV is not able to handle the current situation and (2) the act of watching over a person to detect when the person has become inattentive while watching over an AV. This paper uses "supervising" for the former and "monitoring" for the latter, even when describing work that uses other terminology.

[2] An inattentive RHV is not necessarily a distracted RHV. He or she could be asleep, or so focused on one part of the driving that he or she does not notice something critical happening to the side. So, "inattentive" is used as the most general term for what an RHV should not become, and "distracted" is not a synonym for "inattentive".

[3] "Informing" is the simple act of the Monitor's telling the Notifier that the RHV appears to have become inattentive, while "notification" is the complex act of the Notifier's somehow telling the RHV that he or she needs to be more attentive. Notification is any of a spectrum of acts, ranging from just gently telling the RHV to be more attentive to causing the vehicle to do something that the RHV will surely notice, all with the goal of reengaging the driver to be attentive.

been detected in the RHV.

Even a Monitor is never perfect, both failing to detect an inattentive RHV, in a *false negative (FN)*, and incorrectly detecting inattention in an attentive RHV, in a *false positive (FP)*. Generally, the developers of the Monitor have to trade FNs off with FPs, not being able to eliminate both. An FN is bad because an RHV who has, e.g., fallen asleep is not detected and notified. An FP is bad because the resulting notification contributes to the RHV's perceiving the RMN as crying "wolf!" and then to the RHV's ignoring the RMN. As shown in Section 3.3, the assumption in the literature seems to be that FPs and FNs are equally bad and that recall and precision should be weighted equally. However, this assumption may not be true in some circumstances. *The contribution of this paper is to identify the circumstances in AV operation under which each of FPs and FNs are to be avoided and are to be tolerated. This contribution informs requirements engineering (RE) for the RMN, telling the requirements analyst what data need to be gathered in making the correct tradeoff.*

In the rest of this paper, Section 2 examines the concept of human-centered automation from the aviation industry. The aviation industry successfully dealt with a similar automation shift; so perhaps, similar principles could be applied to AVs as well. Section 3 discusses possible sets of requirements for an RMN and their implications on the tradeoff between FNs and FPs, i.e., between recall and precision, in Monitors. Finally, Section 4 summarizes the paper and describes future work. Due to page limitations, some material that can be found in a full paper [8] was left out. This material includes (1) a discussion of some causes of failure in AVs, (2) an examination of studies involving driving and flying simulations to see how a poorly conceived RMN can have negative effects on RHVs, and (3) a deeper discussion of some related work in the literature.

2 Human-Centered Automation

The actions of aircraft pilots today demonstrate how a vigilant, well-trained human supervisor can provide safety and security in safety-critical domains, but it wasn't always this way. Similar to what is happening today with AVs, aircraft underwent an automation shift starting in the mid 1970s. Functions such as flight path, power control, landing gear, and other subsystems had transformed into fully automated processes. Airplanes were described as completely autonomous by 1991: "current aircraft automation is able to perform nearly all of the continuous control tasks and most of the discrete tasks required to accomplish a mission" [5].

However, as early as 1977, the U.S. House Committee on Science and Technology had said that automation was a major safety concern for the coming decade [15]. The committee's prediction proved to be spot on. A 1990 NTSB report identified 31 of the aviation accidents *involving flight crew* from 1978-1990 were caused by what the report calls "monitoring failures", i.e., aircraft pilots who had become inattentive while they were supposed to be supervising almost completely autonomous aircraft [6]. As a result, NASA suggested that NASA and the FAA adopt a set of human-centered automation (HCA) principles to be used to design the automation on an aircraft [5]. These principles suggest strategies for developing automated systems that help human operators accomplish their responsibilities.

In HCA, automation technology is called on to focus on *helping* aircraft pilots, not replace them. HCA started being applied to aircraft development in the mid 1990s. It kept pilots at the center of responsibility and control. Data from ACRO [1] demonstrate the significant reduction in air incidents in the years after HCA was introduced. Even with HCA, pilots still get bored, leading to lapses in attention. To cope with boredom, pilots engage in secondary tasks such as doing puzzles, talking to colleagues, paying mental games, fidgeting, looking around, and reading training manuals. Despite not tending to the primary task of flying, studies with simulations have shown that pilots who relieved boredom through these activities were less likely than those who did nothing to abdicate responsibility to the automation or to fail to supervise the automation properly [4].

Admittedly, these activities that pilots are allowed and encouraged to do are not realistic in vehicles. A pilot is required to fly with at least one other pilot in the cockpit; a driver does not always have a co-driver to talk to. Also, reading and playing games aren't good options either: a driver requires a faster reaction time in order to avoid hazards, such as pedestrians, cars, objects, etc., which are more abundant on the ground than those in the air. It is unlikely that the deceased in either the Uber or the Tesla incident would not have been killed if the RHV in the incident had been playing games instead of paying attention. *Therefore, to successfully apply HCA to AVs, it will be necessary to discover and invent ways to keep RHVs engaged in ways that allow and encourage very fast response to unexpected events.*

3 Requirements for the RMN

Based on the functionality of an RMN, described in Section 1.1, this section identifies the requirements for the Monitor and the Notifier and explores related work to learn what is feasible for each and what might be traded off between them.

What is learned informs requirements specification for an RMN that is as effective as possible[4].

The subsections of this section describe (1) suggestions for implementations of a Monitor and how to evaluate its effectiveness, (2) suggestions for implementations of a Notifier and how to evaluate its effectiveness, and (3) the tradeoffs that can be made in an effort to achieve the most effective overall RMN, consisting of a Monitor and Notifier working together.

3.1 The Monitor and its Evaluation Existing algorithms for monitoring an AV's RHV use data gathered by various devices that continually observe the RHV and compute predictions about whether the RHV is engaged. For example, Braunagel *et al.* conducted an eye- and head-tracking study of 73 participants using a driving simulator while doing a number of activities, both driving and not driving [7]. Eye-tracking cameras were particularly important to the study: quick eye-movements may indicate that the driver is aware of his or her surroundings, while long gazes too far to the left or right may indicate that the driver is distracted. Other data captured from the cameras include eye-blink frequency and the head's angle and position. Using a multi-class support vector machine algorithm, each participant was predicted as either reading, composing e-mail, watching a movie, or paying attention to driving. One variation of the algorithm predicted with 70% precision, 76% recall, and 77% accuracy. To decide which algorithm, or variation thereof, is better, Braunagel *et al.* say to use accuracy, which weights FNs and FPs equally, as the arbiter.

Thus, the general plan for an RMN is that at regular, frequent intervals of time, its Monitor is examining its input to make a *decision* about whether the RHV is inattentive. If the Monitor determines that the RHV is inattentive, the Monitor informs the RMN's Notifier. Because a notification happens in response to a decision of inattentiveness, and a decision of inattentiveness is expected to be rarer than a decision of attentiveness, inattentiveness is considered the *positive* decision and attentiveness is considered the *negative* decision in the discussion below.

The most common measures used to evaluate the decision-making effectiveness of an AI that functions as a Monitor are the traditional *recall* (R), *precision* (P), *accuracy* (A), the F*-measure*, and its weighted variation [16, 3], $F_\beta = (1 + \beta^2) \times \frac{P \times R}{(\beta^2 \times P) + R}$. In this formula, β is the ratio by which it is desired to weight R more than P, a ratio whose numerator is the cost of an FN and whose denominator is the cost of an FP.

3.2 The Notifier and its Evaluation In the typical current AV that has an RMN, when the Monitor has decided that the RHV is inattentive, the Monitor informs the Notifier to do its job. The typical Notifier then directly notifies the RHV that he or she is inattentive, perhaps activating a speaker to say "Please pay attention!". It is hoped that after such a notification, the RHV begins to pay attention to the driving task. If not, then the Monitor will soon notice that the RHV is still inattentive, and will again ask the Notifier to notify the RHV.

The problem with this simple design for the Notifier is that the effectiveness of a notification that is repeated too often probably begins to deteriorate. The effect of AVs on RHVs described in Sections 3.1 and 3.3 of the full paper [8] suggests that over time, an RHV will begin to treat the notification as background noise and to ignore it.

Therefore, it will be necessary to invent notification techniques whose effectiveness does not deteriorate when notifications are repeated. When automated aircraft designers faced the same effectiveness deterioration problem, with HCA, they found ways to adjust the level of automation in the aircraft itself so that the aircraft's pilots were required to do more in order to fly the aircraft. A pilot who is busy flying the aircraft is naturally engaged and is therefore attentive. Essentially, HCA needs to be applied to the design of an AV and of its RMN and to find ways the gracefully reduce the level of automation of an AV, in order to reengage the RHV.

The design of a graceful reduction in the level of automation of an AV is not easy. Suppose that we have a Level 2 AV, such as the Tesla Model 3 with Autopilot. The difficulty is finding a reduction in automation that is indeed graceful. Just stopping steering or throttling without telling the driver could confuse the driver. Just quietly stopping the lane-centering function could be quite dangerous. Probably, it will be necessary for the vehicle (1) to inform the driver about a *specific* upcoming reduction in automation and (2) to require some form of acknowledgement from the driver, before it actually does the reduction. One possibility is for the vehicle to announce both in sound and in text, "I am shutting off cruise control. You will need to control the throttle. Please confirm by touching the touch screen that you are ready to do so." A number of researchers have proposed and explored ways for an AV to pass control of the AV to the RHV in ways that are not causing the RHV to be momentarily disoriented and at risk for a crash [2, 13, 17, 18, 10]. Of course, it will be necessary to experimentally verify that a proposed notification technique is both graceful enough and that its effectiveness does not degrade with repetition.

Thus, the evaluation of a particular Notifier should consist of conducting an experiment to measure the deterioration of the effectiveness of Notifier's notification technique as a function of the frequency with which the Notifier delivers

[4]Effectiveness of an RMN is defined in Section 3.3, when more of the requirements are understood. Until then, a vernacular understanding suffices.

notifications.

3.3 Tradeoffs in an RMN A Monitor and a Notifier cooperate to build an RMN. Thus, requirements for an RMN, particularly those affecting the RMN's effectiveness, have implications on the requirements for its Monitor and for its Notifier. When satisfaction of a requirement is dependent on a tradeoff, it's useful to have metrics to guide the trading.

An RMN is *most effective* when **(1)** its Monitor has 100% recall, and is thus detecting all instances of RHV inattention, and **(2)** the effectiveness of its Notifier's notifications do not degrade when they are repeated. The danger of too many FNs, i.e., low recall, in the Monitor is that the RHV could, e.g., be asleep but is not notified. The danger of too many FPs, i.e., low precision, in the Monitor is that there will be spurious notifications bothering an attentive RHV, and the RHV could eventually learn to ignore the notifications, leading to the degradation of the effectiveness of the notifier. In practice, FNs and FPs, or recall and precision, have to be traded off [3]. To get fewer FNs and higher recall, an algorithm has to suffer more FPs and lower precision, and vice versa. In fact, there are two extremes:

(1) Achieve 100% recall by always notifying the RHV. This extreme amounts to the RHV's manually driving the vehicle.

(2) Achieve 100% precision by never notifying the RHV. This extreme amounts to having a fully-autonomous vehicle.

For now, the second extreme is not acceptable, as for the foreseeable future, self-driving vehicles make too many mistakes. The first extreme is not much better, as it eliminates the autonomy of an AV. So, the goal for an AV's RMN is for its Monitor to achieve as high a recall as possible without degrading the effectiveness of its Notifier and without defaulting into the RHV's just manually driving the vehicle.

It appears that all 13 items in the literature known to the authors about monitoring algorithms manage the tradeoff only implicitly, by *assuming*, sometimes with no discussion, that FNs and FPs are equally bad, i.e., (1) in the formula for accuracy, the four regions of the space of decisions are weighted equally, and (2) in the formula for F_β, $\beta = 1$ and thus, R and P are weighted equally. A discussion of this literature is found in Section 5.5 of the full paper [8].

However, suppose that the designers of AVs learned from the experiences of aircraft designers' introduction of automation to aircraft cockpits and applied HCA to design Notifiers with notification techniques whose effectiveness does not degrade with repetition. Then, the associated Monitor should be designed with the highest recall possible, even at the cost of a moderately low precision (not to the extreme of always reporting inattention), as is done for AIs that examine medical data to screen patients for diseases particularly when a follow-up test is cheap [9]. So, if a vehicle's response to a prediction of inattentiveness is one that does not start to be ignored in the presence of FPs, the algorithm, or variation thereof, with the highest recall should be chosen. For example, among Braunagel *et al.*'s algorithms, the one with the highest recall achieved only 76% recall. This recall is not very good, because the algorithm is failing to detect 24% of the inattentive spells, leaving the vehicle with no supervision almost one-quarter of the time. It would pay for Braunagel *et al.* to play more with the algorithm to see if tolerating more imprecision can bring recall closer to 100% than 76%.

In deciding the tradeoff for the Monitor, it is essential to compare the costs of an FP and of an FN in the whole RMN.

(1) The effect of an FN is for the system to not notice that the RHV is inattentive. The expected cost of an FN is (1) the probability of an FN, times (2) the probability that an accident will happen when the RHV is not attentive, times (3) the average cost of an accident.

(2) The effect of an FP is for the RHV to be notified of inattentiveness unnecessarily. The expected cost of an FP is (1) the probability of an FP, times (2) the probability that an FP will finally teach the RHV to ignore warnings thus making warnings useless and leaving the driver inattentive after all, times (3) the average cost of an accident.

These two expected costs have to be compared in any situation. Observe that the average cost of an accident is a factor in both expected costs, so the comparison is between two products of probabilities. The ratio of these two products of probabilities can be used as the ratio of the importance of recall to the importance of precision in any situation, and thus as the β in the formula for F_β to evaluate the Monitor.

Berry explains how to obtain and calculate the data, including β, for doing a thorough evaluation of a tool based on its recall, precision, and context [3]. Winkler *et al.* show how to use F_β not only to evaluate the effectiveness of an ML-based tool for classifying natural language statements of requirements, but also how to optimize the tool according to, and in the direction indicated by, the weight β (for recall if $\beta > 1$) and to a degree consistent with the difference between 1 and β [19].

4 Conclusion and Future Work

This paper recounts the circumstances of two fatal accidents involving AVs during which their RHVs failed to maintain the required attentiveness. It shows how HCA helped the aviation industry successfully counteract pilot inattentiveness and suggests ways to do the same with AVs.

The main point of the paper is that if HCA is applied to the design of the Notifier of an RMN to produce a Notifier whose effectiveness in bringing the RHV back to attentiveness does not degrade in the face of too frequent notifications, then the Monitor of the RMN can be safely optimized for fewer FNs, or higher recall, at the cost of more FPs, or lower precision, to obtain a more effective overall RMN. Heretofore, the assumption has been that FPs and FNs are equally bad and that recall and precision should be weighted equally.

Application of HCA to ordinary vehicle owners who have opted to buy AVs at Level 2 or 3, taking on the role of RHVs, will be a challenge. AV RHVs are nowhere as well trained as aircraft pilots, and their emergencies have a much shorter time frame than those of aircraft pilots. It will be necessary to invent notification techniques that are both *sustainably* effective and not so disruptive as to momentarily disorient the RHV. The authors admittedly have difficulty thinking of such notification techniques. However, we are not going to begin to find any such techniques if we don't look for them with the knowledge that they can be used. That said, there is one class of RHVs for which HCA may work, namely professional drivers as for taxis and trucks. These drivers have known how to use walkie-talkies for years without becoming distracted from driving.

Thus, there is a need for future work in the simultaneous design of high recall Monitors and low degradation Notifiers for use in high effectiveness RMN for AVs. The lower the degradation of the Notifier's effectiveness the more FPs, or low precision, can be tolerated in the quest for few FNs, or high recall, in the Monitor.

Acknowledgements

The authors thank Peter van Beek for his comments and a pointer to relevant literature. Daniel Berry's work was supported in part by a Canadian NSERC Discovery Grant, NSERC-RGPIN227055-15. Krzysztof Czarnecki's work was supported in part by a Canadian NSERC Grant, NSERC-RGPIN262100i-12.

References

[1] ACRO Records. Air incidents 1918–2017, 2017. Wikimedia Commons.

[2] S. Baltodano, N. Martelaro, R. Maheshwari, D. Miller, W. Ju, N. Gowda, and S. Sibi. Nudge: Haptic pre-cueing to communicate automotive intent. *Automotive User Interfaces*, 15, 2015.

[3] D. Berry. Evaluation of tools for hairy requirements and software engineering tasks. In *Proceedings of Workshop on Empirical Requirements Engineering (EmpirRE) in IEEE 25th International Requirements Engineering Conference Workshops*, pages 284–291, 2017.

[4] H. Bhana. *Correlating Boredom Proneness With Automation Complacency in Modern Airline Pilots*. PhD thesis, University of North Dakota, Grand Forks, North Dakota, USA, 2009.

[5] C. Billings. Human-centered aircraft automation: A concept and guidelines. Technical Report NASA Technical Memorandum 110381, NASA Ames Research Center, Aug. 1991.

[6] N. T. S. Board. A review of flight crew involved major accidents of u.s. air carriers, 1978 through 1990. Technical Report NTSB/SS-94/01 Notation 6241, National Transportation Safety Board, Jan 1994.

[7] C. Braunagel, E. Kasneci, W. Stolzmann, and W. Rosenstiel. Driver-activity recognition in the context of conditionally autonomous driving. In *IEEE 18th International Conference on Intelligent Transportation Systems*, pages 1652–1657, Sept. 2015.

[8] J. DiMatteo, D. M. Berry, and K. Czarnecki. Requirements for monitoring inattention of the responsible human in an autonomous vehicle: The recall and precision trade-off. Technical report, University of Waterloo, 2020.

[9] W. Koehrsen. Beyond accuracy: Precision and recall: Choosing the right metrics for classification tasks, 2018. Towards Data Science.

[10] T. Lee. Another Tesla driver apparently fell asleep—here's what Tesla could do. *arsTECHNICA*, 2019.

[11] National Transportation Safety Board. Highway preliminary report: Hwy18mh010. Technical report, National Transportation Safety Board, May 2018.

[12] National Transportation Safety Board. Highway preliminary report: Hwy19fh008. Technical report, National Transportation Safety Board, May 2019.

[13] T. Nukarinen, J. Rantala, A. Farooq, and R. Raisamo. Delivering directional haptic cues through eyeglasses and a seat. In *IEEE World Haptics Conference (WHC)*, pages 345–350, 2015.

[14] On-Road Automated Driving (ORAD) Committee. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles, revised standard. Technical Report J3016_201806, SAE International, 2018.

[15] R. Parasuraman, T. Bahri, J. Deaton, and J. Morrison. Theory and design of adaptive automation in aviation systems. Technical Report ADA254595, Defense Technical Information Center, July 1992.

[16] T. Saracevic. Evaluation of evaluation in information retrieval. In *SIGIR Conf. Res. & Devel. Inform. Retrieval (SIGIR)*, pages 138–146, 1995.

[17] R. van der Heiden, S. Iqbal, and C. Janssen. Priming drivers before handover in semi-autonomous cars. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI)*, pages 392–404, 2017.

[18] M. Walch, T. Sieber, P. Hock, M. Baumann, and M. Weber. Towards cooperative driving: Involving the driver in an autonomous vehicle's decision making. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Automotive'UI 16, pages 261–268, 2016.

[19] J. P. Winkler, J. Grönberg, and A. Vogelsang. Optimizing for recall in automatic requirements classification: An empirical study. In *27th IEEE International Requirements Engineering Conference (RE)*, pages 40–50, 2019.