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

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

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

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

ABSTRACT

As shown by recent fatal accidents with partially autonomous vehicles (AVs), the responsible human in the vehicle (RHitV) 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. As shown by aviation's experience and various simulation studies, as the level of automation of an AV increases, the tendency for the RHitV to become inattentive grows. To counteract this tendency, an AV needs to monitor its RHitV for inattention and when inattention is detected, to somehow notify the RHitV to pay attention. The monitoring software needs to tradeoff false positives (FPs) and false negatives (FNs) (or recall and precision) in detecting inattention. FNs (low recall) are bad because they represent not detecting an inattentive RHitV. FPs (low precision) are bad because they lead to the RHitV's being notified frequently, and thus to the RHitV's ignoring notifications as noise, i.e., to degraded effectiveness of notification. The literature shows that most researchers just assume that FPs and FNs (recall and precision) are equally bad 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.

CCS CONCEPTS

• **Software and its engineering** → **Requirements analysis**; • **Computing methodologies** → *Robotic planning*; • **Human-centered computing** → Ubiquitous and mobile computing design and evaluation methods;

KEYWORDS

automation, complacency, engagement, engineering, self-driving vehicle, software

ACM Reference Format:
Johnathan DiMatteo, Daniel M. Berry, and Krzysztof Czarnecki. 2020. Requirements for Monitoring Inattention of the Responsible Human in an Autonomous Vehicle: The Recall and Precision Tradeoff. In *Proceedings of Not Accepted to Any ACM Conference (NONE).* ACM, New York, NY, USA, Article 0, 10 pages. https://doi.org/None

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
NONE, 2020, Nowhere
© 2020 Copyright held by the owner/author(s).
ACM ISBN None.
https://doi.org/None

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 discovered that in 31 of the 37 serious accidents involving U.S. air carriers from 1978–1990, inadequate attention played a major role [8]. Pilots or crew members neglected to govern instrumentation, verify inputs, or communicate caught errors. During the period of review by NTSB, the aviation industry was undergoing significant 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.

- On a dark night in March, 2018, an Uber Technologies, Inc. vehicle was being tested operating fully autonomously when the vehicle suddenly struck and killed a pedestrian who was crossing the street (Figure 1). A preliminary NTSB report of the crash revealed that the supervisor, i.e., the backup driver — responsible for commandeering and then operating the vehicle in times of emergency — was distracted moments before the incident [25]. The supervisor's hands were not on the steering wheel, her eyes were directed downward just before impact, and she was unable to engage the emergency brakes. The report concludes that the vehicle's operating software had apparently recognized the pedestrian as something else.
- 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. The top half of the Tesla was sheared off, and the driver of the Tesla died as a result of the crash (Figure 2). No evasive maneuvers had been executed. As in the Uber incident, a preliminary NTSB report revealed that the Tesla driver was not paying proper attention to his vehicle and the road conditions [26]. Sensors

Figure 1: Location of the crash, showing paths of pedestrian in orange and the Uber vehicle in green [25].

Figure 2: Post-crash view of the Tesla Model 3 [26].

in the vehicle showed that the driver's 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. Part of the problem is that moving more and more of the vehicle's operation into the vehicle's software causes the requirements of the vehicle to change. The now-autonomous vehicle (AV) must gather data from its surroundings that it had never needed before. The role of the driver has changed too. Previously, driving a vehicle required constant attention and action on the part of the driver. Now, a driver may not be needed at all.

Table 1: Automation levels defined by the Society of Automotive Engineers (SAE) J3016 [30]

Level	RHitV's Duty	Vehicle's Degree of Automation
0	drives	totally manual
1	drives	assists driver with some functions[1]
2	drives	automates some functions[2]
3	supervises	automates many functions[3]
4	nothing	automates all functions needed in specific conditions
5	nothing	automates all functions needed in all conditions

The increasing autonomy is shifting the driving responsibility from the driver to the AV's 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 the vehicle (RHitV)* that is supposed to be driving or supervising[4] the AV. A RHitV is *supervising* a vehicle when the RHitV 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 six autonomy levels (Table 1), with which to classify the degree of automomy of a vehicle according to (1) the degree of human attention required and (2) the degree of automation of the vehicle [5]. This paper focuses on two particular levels of autonomy:

Level 2 in which there is a RHitV tasked with driving the vehicle even though some of the functions of the vehicle are automated; that is, the RHitV 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

[1] Vehicle does ((steering *or* speed control) and *partial* perception). Driver does either steering or speed control, and is thus engaged continuously.
[2] Vehicle does ((steering *and* speed control) and *partial* perception). Driver does neither steering nor speed control and only completes perception when the vehicle cannot, and thus is engaged only rarely.
[3] Vehicle does ((steering and speed control) and *full* perception). Driver does neither steering, speed control, nor any perception, but must be receptive at all times (1) to anything that might be unusual and (2) to requests from vehicle, to take over driving.
[4] 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 reserves "monitoring" for the latter. This wording is maintained even when describing work that uses different terminology. So do not be surprised to find that someone else's monitoring is described as "supervising" in this paper.
[5] There are subtleties of the SAE standard that are abstracted away in this table. The subtleties affect the details and weights of any specific tradeoff, but they do not affect the *existence* of the tradeoff.

Level 3 in which there is a RHitV tasked with supervising the vehicle even as the partially autonomous vehicle is doing many functions; that is, while the RHitV 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 to 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 RHitV of a vehicle is (1) a driver or (2) a supervisor, and what a RHitV does, *piloting*[6], is (1) driving or (2) supervising.

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

- the *Monitor*, an AI that somehow *monitors*[7] the RHitV for signs of *inattention*, and at any time that the Monitor *detects* that the RHitV is inattentive[8], it *informs* the Notifier to do its job, and
- the *Notifier*, when informed by the Monitor, somehow *notifies* the AV[9], the RHitV, or both, that signs of inattention have been detected in the RHitV.

Even a Monitor is never perfect, both failing to detect an inattentive RHitV, in a *false negative (FN)*, and incorrectly detecting inattention in an attentive RHitV, in a *false positive (FP)*. Generally, the developers of the Monitor have to trade FNs off with FPs, not being able to eliminate both. A FN is bad because a RHitV who has, e.g., fallen asleep is not detected and notified. A FP is bad because the resulting notification contributes to the RHitV's perceiving the RiMN as crying "wolf!" and then to the RHitV's ignoring the RiMN. As shown in Section 5.5, 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 RiMN, telling the requirements analyst what data need to be gathered in making the correct tradeoff.*

In the rest of this paper, Section 2 discusses some causes of failure in AVs. Section 3 examines studies involving driving and flying simulations to see how a poorly conceived RiMN can have negative effects on human RHitVs. Section 4 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 5 discusses possible sets of requirements for RiMN and their implications on the tradeoff between FNs and FPs, i.e., between recall and precision, in Monitors. Finally, Section 6 summarizes the paper and describes future work.

[6] Indeed, the tasks of an aircraft pilot can be described as a mixture of driving and supervising.

[7] In this paper, for each noun ***n***, the verb of what ***n*** does is the verb that shares a root with ***n***, and vice versa.

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

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

2 FAILURES IN AUTONOMOUS VEHICLES

New technologies, often based on AI powered by ML, are enabling the development of autonomous systems in places never thought possible. However, these systems are not free of failures, and failures can result in the system's not satisfying safety requirements, particularly in safety-critical systems. To provide the lowest risk possible, humans must supervise an autonomous system and respond to its failures. Understanding human factors and the limitations of autonomous systems allows designing an AV and its RiMN to work together to achieve the required level of safety.

An AV must deal with unpredictable surroundings. An AV is not isolated, it must interact with humans: pedestrians, cyclists, and drivers of other vehicles on the road. For example, a human driver could attempt to cut in front of an AV, anticipating that the AV will not respond aggressively. If the AI operating the AV cannot cope with some unpredictability, then there must be a human present in the AV supervising the AV for unpredictable events that the AV cannot handle, ready to take over as the AV's driver.

Some faults present in AVs today are unavoidable even in a Level 5 AV. Unlikely events, such as random bit flips caused by the sun's radiation, can erroneously effect software calculations and cause the AV to behave incorrectly. Hardware component and sensors can malfunction, e.g., as a result of degradation in cold or corrosive weather [16]. The solution for these kinds of faults is hardware fault tolerance, including redundant hardware [2].

There will always be malicious actors trying to exploit security vulnerabilities in AVs. Since an AV is controlled by software, whoever gains control over the AV's software can commandeer the AV, possibly driving it remotely. The solution for these kinds of faults is having programmed security into the AV from the beginning of its development [14].

The AI currently in AVs is error prone, as demonstrated by the recent Uber test vehicle and Tesla Model 3 incidents. Neither the Uber test vehicle nor the Tesla Model 3 had a RiMN working to prevent the fatal lapse of attention on the part of vehicle's RHitV.

The Tesla Model 3 with Autopilot engaged is properly classified as Level 2. While the Uber test vehicle appears to be classified as Level 3, it was in fact a prototype for Uber's planned Robo Taxi, which is targeted to be released as a Level 4 AV. Every prototype AV, regardless of its target level, requires the presence of a RHitV, called a "safety driver", supervising the vehicle in the fullest sense of the word. While the goal of all AV manufacturers is to produce a Level 4 AV, to date, no one has produced a Level 4 or 5 AV. Moreover, correctly, no one claims to have produced one. Therefore, everyone understands that for even the most automated AV available today, a responsible human supervisor is needed in the AV, ready to take over the driving at any time.

After the incident, Tesla announced: "data shows that, when used properly with an attentive driver who is prepared to take over at all times, drivers supported by Autopilot are safer than those

operating without assistance" [35]. Despite the misleading name, Tesla's Autopilot requires an attentive driver who is prepared to take over the full role of driving at any time. Expecting the driver to remain attentive after he or she has engaged a feature that in fact makes driving require *less* attention invites mistakes, as evidenced by the incident. That the feature is named "Autopilot" enhances this invitation. Tesla's statement about proper use of Autopilot is irrelevant when a driver in fact accepts the invitation and becomes inattentive.

More concerning is that traditional RE and safety principles may be ignored as car manufacturers feel pressure to innovate and design the most autonomous system as possible, instead of the safest system possible. Also, consumer overconfidence will present a major challenge, especially in a Level 3 AV, which cannot perform independently in all conditions. As discussed in Section 3, this overconfidence contributes to the tendency of RHitVs, who are often consumers who bought the vehicles that they are piloting, to lapse into inattention.

3 EFFECTS OF VEHICLE AUTONOMY

As evidenced by the Tesla incident, requiring a RHitV to pay attention doesn't mean that the RHitV always will. The RHitV is affected by an AV in ways that the designers of the AV never intended. The resulting effects can make it very difficult for a RHitV to supervise the AV effectively.

3.1 Active and Passive Fatigue

In general, piloting requires effort by the RHitV, and naturally, effort causes fatigue. Automation research has discovered two different types of *fatigue*, namely *active* and *passive*, that can impact a RHitV after long periods of piloting [19]. Active fatigue happens when a driver spends physical energy on the driving task, such as steering, braking, and scanning the road for hazards. Passive fatigue happens when a supervisor, who is not actually performing the driving task, nevertheless becomes fatigued by the mental effort required to remain attentive and focused, to perform the supervising task. Passive fatigue is heightened when there is more information for the supervisor to focus on. Although the symptoms of active fatigue are physical exhaustion, stress, and heightened coping effort, the symptoms of passive fatigue are more subtle: mental fatigue, a decline in task engagement, and infrequent use of controls.

Saxby *et al.* conducted studies to demonstrate that the two types of fatigue can cause different performance effects on the user and user safety [34]. They tested 168 participants for (1) fatigue effects on vehicle control and (2) alertness in a driving simulation. Each participant was told to pilot a vehicle simulator for 30 minutes For this piloting, the participants were split into two groups. Each participant in the first group drove a simulated vehicle in an active setting on a curved road, with winds gusting throughout the 30 minute period, requiring continual correctional steering on the participant's part. Each participant in the second group supervised a fully autonomous vehicle in a passive setting for the 30 minute period. Immediately after the 30-minute piloting, each participant of either group drove a simulated vehicle for four minutes and was told to anticipate an emergency event (Figure 3). The data collected for each participant during the 4-minute driving were (1) whether or not he or she crashed during the anticipated emergency event and (2) the speed of his or her braking and steering response to the emergency event.

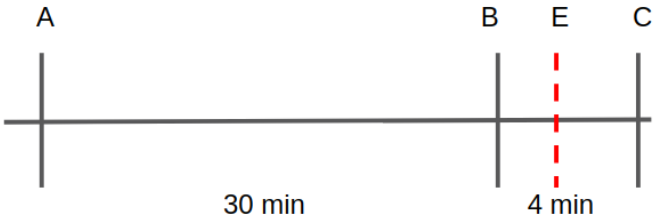

Figure 3: Each participant piloted for 34 minutes total, with the first 30 minutes (A to B) either supervising a fully autonomous vehicle or driving a fully manual vehicle on a curved road with wind gusts [34]. The last 4 minutes was a performance test (B to C) driving a fully manual vehicle. Each participant was told to expect an emergency event (E) sometime during the performance test.

Despite having to endure 30 minutes of physical effort in steering and controlling the simulated vehicle just beforehand, participants in the active setting performed much better in the emergency event than those in the passive setting, *even with the forewarning of the impending emergency event!* The participants of the second group crashed their simulated vehicle more often than did the participants of the first group. Although both types of fatigue caused tiredness, aversion to effort, and loss of task engagement, only passive fatigue reduced alertness. The results suggest that a supervisor spends significant mental energy performing tasks such as staying alert and absorbing information.

3.2 Malleable Attentional Resources Theory

The typical explanation given for a person's deteriorated driving performance after a long period of supervising an AV is mental fatigue. However, Solís-Marcos *et al.* noticed the same deterioration even after only a few minutes of supervising, too soon for any kind of fatigue to have set in [37]. They believed that this deterioration is explained also by Young and Stanton's Malleable Attentional Resources Theory (MART), which attempts to explain why people lose focus so quickly [41]. According to MART, attentional resources shrink to accommodate any demand reduction. When a task becomes easier, people are not inclined to spend the extra time and energy performing some other task closely related to the task at hand. In the case of AVs, if the automation level increases, people in the AV do not use the freed up time to look more carefully at traffic, plan ahead, or seek potential hazards. Instead, they spend only the resources required by the supervising task. MART suggests that future AV designers should employ their technology in pilot-support systems rather than in automation to replace the driver, supporting human-centered automation principles. Otherwise, users will decrease mental effort on the driving task as it becomes easier, ignoring the need to stay vigilant for safety precautions.

Matthews and Desmond conducted a study with a simulated partially autonomous vehicle [24]. First, each subject drove the

vehicle for a lengthy enough period for fatigue to set in. Then, each subject drove a short track that demanded high performance and concentration. The track for half of the subjects was straight and the track for the other half was curved. The straight-track subjects committed significantly more heading errors than did the curved-track subjects. Matthews and Desmond believed that the straight-track subjects underestimated the difficulty of their driving task and withdrew necessary focus. These subjects suffered not only passive fatigue, but the effect of cognitive underload caused by partial automation, as predicted by MART.

MART explains also the apparent loss of attention in the Tesla incident, in which Autopilot was active for a mere 10 seconds before the impact. Other studies have shown that aircraft pilots are effected by MART as well. In a study by Casner *et al.*, sixteen pilots were asked to fly in a Boeing 747 simulator [13]. As the simulated flight progressed, automation levels were varied, and anomalies were introduced at random points in the flight. The pilots were told beforehand that anomalies would happen, so they should always try to pay attention. Despite this warning, only one pilot was able to complete the test without making a mistake, and the rest exhibited significant mind-wandering. In most cases, the pilots could detect what was wrong, but didn't respond as well as they should have. The higher the automation level, the fewer attentional resources were spent on the task, and the worse the errors were. Thus, even though cockpit automation provides pilots more time to think, it may actually encourage pilots to invest only part of this free time in thinking flight-related thoughts.

3.3 Overconfidence

Overconfidence may not be caused directly by supervising itself, but it does have a significant impact on supervising performance. AI technology is exciting but the limitations are not well understood by the general public. Vehicle manufacturers feel enormous pressure to market their products as more autonomous or cutting-edge than other competitors, creating a tension between the need to innovate and the need for safety. As a result of popular fiction about computers and software in entertainment media, the public tends to believe that the capabilities of AI systems are greater than they actually are, that AI systems are even *truly adaptively intelligent*. Overconfidence on a driver's part naturally leads to complacency, which causes drivers to be unaware of dangers around them. Complacency is especially dangerous in safety-critical systems. As happened in the Uber and Tesla incidents, people are willingly putting their lives — and the lives of people around them — at risk by trusting AVs to work flawlessly. The public needs to understand how automation works to realize that AI systems are not foolproof.

4 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

Figure 4: The cockpit of a Boeing 747 airplane has dozens of dials, screens, and gauges that require supervising.

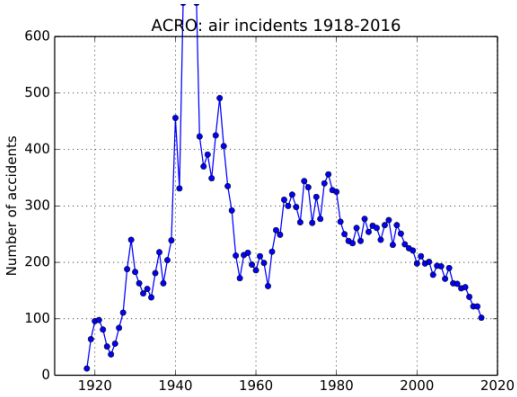

Figure 5: Statistics of air accident incidents 1918-2017, from ACRO records [1].

nearly all of the continuous control tasks and most of the discrete tasks required to accomplish a mission" [7]. The result was that inside the cockpit of a Boeing 747, there are dozens of dials, gauges, and screens that are operating automatically (Figure 4).

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 [31]. 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 [8]. Even before this NTSB report was finally officially published in 1994, NASA and the FAA had identified the growing challenges of automating aircraft. Consequently, a 1991 NASA report observes that

> Several aircraft accidents and a larger number of incidents have been associated with, and in some cases appear to have been caused by, aircraft automation or, mote accurately, by the interaction between the human operators and the automation in the aircraft.

and suggests that NASA and the FAA adopt a set of human-centered automation (HCA) principles to be used to design the automation

on an aircraft [7]. 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. The NASA and FAA report said that "the quality and effectiveness of the pilot-automation system is a function to the degree of which the combined system takes advantages of the strengths and compensates for the weaknesses of both elements" [7]. Humans have a number of abilities that they do far better than does any AI now and for the foreseeable future:

(1) detecting signals or information in the presence of noise,
(2) reasoning efficiently and effectively in the face of uncertainty, and
(3) abstracting and conceptually organizing.

These abilities make humans irreplaceable.

New research in deep learning is attempting to replicate human qualities, but there is still a long way to go, if possible at all. Therefore,

- any function that humans do best should be left to the human pilot at hand, and
- if during autonomous operation, the vehicle needs assistance that can best be rendered by humans, the human pilot should be called on, even in a non-emergency, if for no other reason than to keep the human pilot engaged.

Regardless, a safety-critical domain is not an area in which to test innovative technology if there is any chance that safety will be impacted.

HCA started being applied to aircraft development in the mid 1990s. It kept pilots at the center of responsibility and control. Statistics demonstrate the significant reduction in air incidents in the years after HCA was introduced (Figure 5). Even with HCA, pilots still get bored, leading to lapses in attention. Pilots are human after all. 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 [6].

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 RHitV 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 vehicle RHitVs engaged in ways that allow and encourage very fast response to unexpected events.*

5 REQUIREMENTS FOR THE RIMN

Recall the functionality of the RiMN:
The RiMN consists of two communicating parts:

- the *Monitor*, an AI that somehow *monitors*[10] the RHitV for signs of *inattention*, and at any time that the Monitor *detects* that the RHitV is inattentive, it *informs* the Notifier to do its job, and
- the *Notifier*, when informed by the Monitor, somehow *notifies* the AV, the RHitV, or both, that signs of inattention have been detected in the RHitV.

The goal of this section is to begin to flesh out the details of the requirements for each of the Monitor and the Notifier. Therefore, this section explores related work to learn what is feasible for each of the Monitor and the Notifier and what might be traded off between them. What is learned informs requirements specification for a RiMN that is as effective as possible[11].

The subsections of this section describe (1) past suggestions for implementations of a Monitor, (2) how to evaluate the effectiveness of a Monitor, (3) past suggestions for implementations of a Notifier, (4) how to evaluate the effectiveness of a Notifier, and (5) the tradeoffs that can be made in an effort to achieve the most effective overall RiMN, consisting of a Monitor and Notifier working together.

5.1 The Monitor

Many have developed and evaluated algorithms for monitoring an AV's RHitV. The algorithms use data gathered by various devices that continually observe the RHitV and compute predictions about whether the RHitV is engaged. These decisions are computed at very frequent intervals to minimize the amount of time between when a RHitV becomes inattentive and when he or she is notified. This subsection considers two such algorithms and their evaluations.

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 [10]. 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 the correct activity with 50% precision, 57% recall, and 53% accuracy[12]. This variation performs significantly worse than in a controlled lab environment. So, it is not suitable. Another variation of the algorithm predicted with 70% precision, 76% recall, and 77% accuracy. This variation is more suitable than the previously mentioned one. To decide which algorithm, or variation thereof, is better, Braunagel *et al.* used accuracy (called "ACC" in their paper), which weights FNs and FPs equally, as the arbiter. That is, the variation with the highest ACC measure is considered the best to use.

To monitor whether a RHitV is capable of providing adequate supervision for his or her AV, Fridman *et al.* propose using biometric sensors and cameras [17]. The sensors report the RHitV's heart

[10] In this paper, for each noun ***n***, the verb of what ***n*** does is the verb that shares a root with ***n***, and vice versa.
[11] Effectiveness of a RiMN is defined in Section 5.5, when more of the requirements are understood. Until then, a vernacular understanding suffices.
[12] See Section 5.2 for definitions of these standard measures.

rate and skin conductance, and the cameras are focused on the RHitV's face, eyes, and torso. Fridman *et al.* use image processing to recognize the RHitV's facial expressions, eye movements, and body posture. They use a deep learning algorithm to learn from the RHitV's (1) heart rate, (2) skin conductance, (3) facial expressions, (4) eye movements, (5) posture, and (6) current health the RHitV's current state and whether he or she is attentive enough to be adequately supervising the AV that he or she is piloting. For example,

- a forward-leaning face and closed eyes might indicate that the RHitV is asleep
- a high heart rate might indicate that the RHitV is stressed or anxious, while a low heart rate might cause fatigue or dizziness in the RHitV,
- high skin conductance might indicate that the RHitV is physiologically aroused from external or internal stimuli.

Fridman *et al.* try several variations of the deep learning algorithm in an effort to improve the accuracy of its prediction of whether the RHitV is capable of providing adequate supervision. Generally, the more data the algorithm uses, the better the prediction. By increasing the number of data items used by the algorithm, the accuracy of the algorithm's prediction of the driver's state can be improved from about 62% to about 92%. However using more data increases the time to the next prediction and thus decreases the frequency of estimates. A crash could occur in between two predictions that are too far apart.

Thus, the general plan for a RiMN is that at regular, frequent intervals of time, its Monitor is examining its input to make a *decision* about whether the RHitV is inattentive. If the Monitor determines that the RHitV is inattentive, the Monitor informs the RiMN's Notifier. The Monitor is making decisions at only regular intervals of times. Therefore, each of inattentiveness and attentiveness is an instantaneous phenomenon. 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.

5.2 Evaluation of Monitors

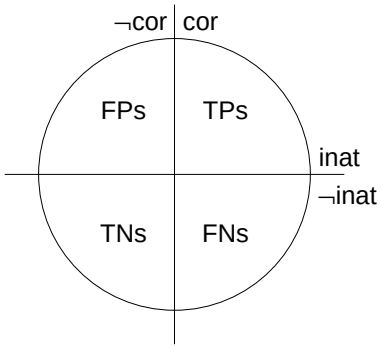

Figure 6: The Universe of a Monitor

The most common measures used to evaluate the decision-making effectiveness of an AI that functions as a Monitor are *recall* (R), *precision* (P), *accuracy* (A), and the *F-measure*. To define these measures precisely, it is necessary to consider in Figure 6 the universe of a Monitor [5]. The circle represents the space of all decisions of the Monitor, correct or not. The space can be partitioned by two independent axes,

(1) one separating the decisions of **inat**tentiveness, inat, from those of attentiveness, i.e., **non-inat**tentiveness, ¬inat, and
(2) one separating the decisions that are correct, cor, from those that are not correct, ¬cor.

These two partitions create four regions in the space, consisting of

- the true positives, TPs, the decisions of inattentiveness that are correct, i.e., the RHitV *is, in fact,* inattentive,
- the false negatives, FNs, the decisions of attentiveness that are incorrect, i.e., the RHitV *is, in fact,* inattentive,
- the true negatives, TNs, the decisions of attentiveness that are correct, i.e., the RHitV *is, in fact,* attentive,
- the false positives, FPs, the decisions of inattentiveness that are incorrect, i.e., the RHitV *is, in fact,* attentive.

With these subspaces, it is possible to give precise definitions of recall, R, and precision, P:

$$R = \frac{|\text{inat} \cap \text{cor}|}{|\text{cor}|} = \frac{|\text{TPs}|}{|\text{TPs}| + |\text{FNs}|} \tag{1}$$

$$P = \frac{|\text{inat} \cap \text{cor}|}{|\text{inat}|} = \frac{|\text{TPs}|}{|\text{FPs}| + |\text{TPs}|} \tag{2}$$

Accuracy, A is defined:

$$A = \frac{|\text{inat} \cap \text{cor}| + |\neg\text{inat} \cap \neg\text{cor}|}{|\text{cor} \cup \neg\text{cor}|} \tag{3}$$

$$= \frac{|\text{TPs}| + |\text{TNs}|}{|\text{TPs}| + |\text{FPs}| + |\text{TNs}| + |\text{FNs}|} \tag{4}$$

Accuracy is the fraction of the entire space that is classified accurately, whether as a TP or as a TN. A decision of inattentiveness or of attentiveness that is correct is considered classified accurately.

The composite of recall and precision that is often called "correctness" is the F-measure:

$$F = 2 \times \frac{P \times R}{P + R}\,, \tag{5}$$

the harmonic mean of recall, R, and precision, P.

Each of accuracy and the F-measure weights all regions of the space, and thus recall and precision, equally. However, there may be situations in which the expected cost of a FN, a failure to detect inattention, is higher than the expected cost of a FP, a spurious detection of inattention. Then, FNs should be weighted more than FPs, and as FNs and FPs are weighted so are recall and precision, respectively. For situations in which recall and precision are not equally important, there is a weighted version of the F-measure,

$$F_\beta \;=\; (1 + \beta^2) \times \frac{P \times R}{(\beta^2 \times P) + R}\,, \tag{6}$$

called the F_β-measure.

Note that the simple F-measure is F_1. Note also that the formula for F_β ends up multiplying P by the dominating β^2 in both the numerator and the denominator. Thus, as β grows, very quickly F_β approaches R, and P becomes irrelevant in computing F_β. When β is as little as 5, and P is large enough relative to R, P is already essentially irrelevant.

In these formulae, β is the ratio by which it is desired to weight R more than P, a ratio whose

numerator is the cost of a FN and
denominator is the cost of a FP.

Section 5.5 explains how to compute these costs for any particular RiMN.

5.3 The Notifier

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

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 RHitVs described in Sections 3.1–3.3 suggests that over time, a RHitV 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 they 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 RiMN.

In 2018, the SAE released the J3016 standard taxonomy related to AVs [30]. The standard defines the Operational Design Domain (ODD) — the specific conditions under which a system can operate safely — for automation systems in vehicles. Examples of conditions include geographical areas, road types, traffic conditions, and vehicle speeds. To take this idea further, Colwell *et al.* propose a Restricted Operational Domain (ROD) specified at the requirements level [15]. The ROD is essentially a dynamic version of the ODD, restricting the capabilities of the system at runtime in a graceful manner. As sensors degrade or functionality becomes limited, the ROD adapts and becomes smaller. When the ROD is smaller, human intervention is required more frequently, and the human naturally remains attentive.

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 center 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 RHitV in ways that are not causing the RHitV to be momentarily disoriented and at risk for a crash [3, 23, 27, 38, 40]. 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.

5.4 Evaluation of Notifiers

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 notifications.

5.5 Tradeoffs in a RiMN

A Monitor and a Notifier cooperate to build a RiMN. Thus, requirements for a RiMN, particularly those affecting the RiMN'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.

A RiMN is *most effective* when

(1) its Monitor has 100% recall, and is thus detecting all instances of RHitV 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 RHitV could 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 RHitV, and the RHitV could eventually learn to ignore the notifications, leading to the degredation of the effectiveness of the notifier.

In practice, FNs and FPs, or recall and precision, have to be traded off [5]. 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:

- Achieve 100% recall by always notifying the RHitV. This extreme amounts to the RHitV's manually driving the vehicle.
- Achieve 100% precision by never notifying the RHitV. 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 RiMN is for its Monitor to achieve as high a recall as possible without degrading the effectiveness of its Notifier and without defaulting into the RHitV's just manually driving the vehicle.

It appears that all the literature known to the authors about monitoring algorithms manage the tradeoff only implicitly, by *assuming* 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. Certainly Braunagel *et al.* and Fridman *et al.*, whose work is described in Section 5.1, do so. In addition,

- each of the references [4, 9, 11, 18, 20, 22, 28, 29, 32, 33, 36, 39] evaluates the algorithms it describes using accuracy with the standard formula, which weights FPs and FNs equally;

- each of the references [9, 22] evaluates the algorithms it describes using F_1, which weights recall and precision, and thus, FPs and FNs equally;
- each of the references [12, 20, 22, 32] evaluates the algorithms it describes using recall and precision, and it has no mention of any weighting between them or between FPs and FNs; and
- the reference [22] evaluates the algorithms it describes using also the κ statistic and ROC curves, and it has no mention of any weighting between FPs and FNs.

Note that some references appear more than once in the above list, because each uses more than one measure to evaluate its algorithms. Only one of these works shows any awareness that FNs and FPs may not be equally bad. Ohn-Bar *et al.*, while using the region-balancing accuracy measure to evaluate their algorithms, nevertheless, try specifically to minimize FPs [29].

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 (only moderately low, so that the extreme of always reporting inattention is not used). 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. A recall of 76% is not very good, because this recall means that 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%.

Section 5.4 discusses the issues in achieving a Notifier whose effectiveness does not degrade in the face of many FPs from the Monitor.

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

- The effect of a FN is for the system to not notice that the RHitV is inattentive. The expected cost of a FN is (1) the probability of a FN, times (2) the probability that an accident will happen when the RHitV is not attentive, times (3) the average cost of an accident.
- The effect of a FP is for the RHitV to be notified of inattentiveness unnecessarily. The expected cost of a FP is (1) the probability of a FP, times (2) the probability that a FP will finally teach the RHitV 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 the expected cost of a FN to the expected cost of a FP 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.

Apparently, those applying AI to examining data to screen patients for diseases have learned to tolerate low precision to achieve a recall of close to 100%, particularly when the cost of a more precise follow-up test is low [21].

6 CONCLUSION AND FUTURE WORK

This paper recounts the circumstances of two fatal accidents involving AVs during which their RHitVs failed to maintain the required attentiveness. It discusses some causes of failure in AVs and examines studies involving driving and flying simulations to understand how and why increased automation in AVs leads to greater inattentiveness on the part of the AVs' RHitVs and to see how a poorly conceived RiMN can exacerbate the inattentiveness problem. 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 a RiMN to produce a Notifier whose effectiveness in bringing the RHitV back to attentiveness does not degrade in the face of too frequent notifications, then the Monitor of the RiMN 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 RiMN. 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 RHitVs, will be a challenge. AV RHitVs are nowhere as well trained as aircraft pilots, and their emergencies have a much shorter time frame than those of pilots. It will be necessary to invent notification techniques that are both *sustainably* effective and not so disruptive as to momentarily disorient the RHitV. 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 RHitVs 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 RiMN for AVs. The lower the degredation 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.

ACKNOWLEDGMENTS

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. 2017. Air Incidents 1918–2017. https://commons.wikimedia.org/wiki/File:ACRO_incidents.svg Wikimedia Commons.
[2] A. Aviziens. 1976. Fault-Tolerant Systems. *IEEE Trans. Comput.* 25, 12 (dec 1976), 1304–1312.

[3] S. Baltodano, N. Martelaro, R. Maheshwari, D. Miller, W. Ju, N. Gowda, and S. Sibi. 2015. Nudge: Haptic Pre-Cueing to Communicate Automotive Intent. *Automotive User Interfaces* 15 (2015). https://www.auto-ui.org/15/p/workshops/2/6_Nudge-%20Haptic%20Pre-Cueing%20to%20Communicate%20Automotive%20Intent_Gowda.pdf

[4] A. Banerjee, S. Datta, A. Konar, D.N. Tibarewala, and J. Ramadoss. 2014. Cognitive Activity Recognition Based on Electrooculogram Analysis. In *Advanced Computing, Networking and Informatics (ICACNI) – Volume 1 (SIST, volume 27)*. Springer International Publishing, Cham, 637–644.

[5] D.M. Berry. 2017. 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*. 284–291.

[6] H. Bhana. 2009. *Correlating Boredom Proneness With Automation Complacency in Modern Airline Pilots*. Ph.D. Dissertation. University of North Dakota, Grand Forks, North Dakota, USA. https://commons.und.edu/theses/371

[7] C. Billings. 1991. *Human-Centered Aircraft Automation: A Concept and Guidelines*. Technical Report NASA Technical Memorandum 110381. NASA Ames Research Center. https://ntrs.nasa.gov/search.jsp?R=19910022821

[8] National Transportation Safety Board. 1994. *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. https://apps.dtic.mil/dtic/tr/fulltext/u2/a532150.pdf

[9] C. Braunagel, D. Geisler, W. Rosenstiel, and E. Kasneci. 2017. Online Recognition of Driver-Activity Based on Visual Scanpath Classification. *IEEE Intelligent Transportation Systems Magazine* 9, 4 (2017), 23–36.

[10] C. Braunagel, E. Kasneci, W. Stolzmann, and W. Rosenstiel. 2015. Driver-Activity Recognition in the Context of Conditionally Autonomous Driving. In *IEEE 18th International Conference on Intelligent Transportation Systems*. 1652–1657.

[11] C. Braunagel, W. Stolzmann, E. Kasneci, T.C. Kübler, W. Fuhl, and W. Rosenstiel. 2015. Exploiting the Potential of Eye Movements Analysis in the Driving Context. In *15. Internationales Stuttgarter Symposium*, M. Bargende, H.-Ch. Reuss, and J. Wiedemann (Eds.). 1093–1105.

[12] A. Bulling, J.A. Ward, H. Gellersen, and G. Troster. 2011. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 4 (2011), 741–753.

[13] S.M. Casner, R.W. Geven, M.P. Recker, and J.W. Schooler. 2014. The Retention of Manual Flying Skills in the Automated Cockpit. *Human Factors* 56, 8 (2014), 1506–1516.

[14] S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham, S. Savage, K. Koscher, A. Czeskis, F. Roesner, and T. Kohno. 2011. Comprehensive Experimental Analyses of Automotive Attack Surfaces. In *Proceedings of the 20th USENIX Conference on Security (SEC)*. 6–21. http://www.autosec.org/pubs/cars-usenixsec2011.pdf

[15] I. Colwell, B. Phan, S. Saleem, R. Salay, and K. Czarnecki. 2018. An Automated Vehicle Safety Concept Based on Runtime Restriction of the Operational Design Domain. In *IEEE Intelligent Vehicles Symposium (IV)*. 1910–1917. https://doi.org/10.1109/IVS.2018.8500530

[16] K. Czarnecki. 2018. Requirements Engineering in the Age of Societal-Scale Cyber-Physical Systems: The Case of Automated Driving. In *IEEE 26th International Requirements Engineering Conference*. 3–4.

[17] L. Fridman, D.E. Brown, M. Glazer, W. Angell, S. Dodd, B. Jenik, J. Terwilliger, A. Patsekin, J. Kindelsberger, L. Ding, S. Seaman, A. Mehler, A. Sipperley, A. Pettinato, B.D. Seppelt, L. Angell, B. Mehler, and B. Reimer. 2019. MIT Advanced Vehicle Technology Study: Large-Scale Naturalistic Driving Study of Driver Behavior and Interaction With Automation. *IEEE Access* 7 (2019), 102021–102038. https://doi.org/10.1109/ACCESS.2019.2926040

[18] L. Fridman, P. Langhans, J. Lee, and B. Reimer. 2016. Driver Gaze Region Estimation without Use of Eye Movement. *IEEE Intelligent Systems* 31, 3 (2016), 49–56.

[19] P.A. Hancock and P.A. Desmond. 2001. Active and Passive Fatigue States. In *Human Factors in Transportation. Stress, Workload, and Fatigue*, P.A. Hancock and P.A. Desmond (Eds.). Lawrence Erlbaum Associates, Mahwah, NJ, USA, 455–465.

[20] S. Ishimaru, K. Kunze, K. Kise, J. Weppner, A. Dengel, P. Lukowicz, and A. Bulling. 2014. In the Blink of an Eye: Combining Head Motion and Eye Blink Frequency for Activity Recognition with Google Glass. In *Proceedings of the 5th Augmented Human International Conference (AH)*. 15:1–15:4.

[21] W. Koehrsen. 2018. Beyond Accuracy: Precision and Recall: Choosing the right metrics for classification tasks. https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c Towards Data Science.

[22] O.D. Lara and M.A. Labrador. 2013. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys Tutorials* 15, 3 (2013), 1192–1209.

[23] T.B. Lee. 2019. Another Tesla Driver Apparently Fell Asleep—Here's What Tesla Could Do. *arsTECHNICA* (2019). https://arstechnica.com/tech-policy/2019/09/how-tesla-could-fix-its-sleeping-driver-problem/

[24] G. Matthews and P.A Desmond. 2002. Task-Induced Fatigue States and Simulated Driving Performance. *The Quarterly Journal of Experimental Psychology: Section A* 55, 2 (2002), 659–686.

[25] National Transportation Safety Board. 2018. *Highway Preliminary Report: HWY18MH010*. Technical Report. National Transportation Safety Board. https://www.ntsb.gov/investigations/AccidentReports/Pages/HWY18MH010-prelim.aspx

[26] National Transportation Safety Board. 2019. *Highway Preliminary Report: HWY19FH008*. Technical Report. National Transportation Safety Board. https://www.ntsb.gov/investigations/accidentreports/pages/hwy19fh008-preliminary-report.aspx

[27] T. Nukarinen, J. Rantala, A. Farooq, and R. Raisamo. 2015. Delivering Directional Haptic Cues THrough Eyeglasses and a Seat. In *IEEE World Haptics Conference (WHC)*. 345–350.

[28] E. Ohn-Bar, S. Martin, A. Tawari, and M.M. Trivedi. 2014. Head, Eye, and Hand Patterns for Driver Activity Recognition. In *22nd International Conference on Pattern Recognition (ICPR)*. 660–665.

[29] E. Ohn-Bar and M.M. Trivedi. 2014. Beyond Just Keeping Hands on the Wheel: Towards Visual Interpretation of Driver Hand Motion Patterns. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. 1245–1250.

[30] On-Road Automated Driving (ORAD) Committee. 2018. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, Revised Standard*. Technical Report J3016_201806. SAE International. https://www.sae.org/content/j3016_201806

[31] R. Parasuraman, T. Bahri, J.E. Deaton, and J.G. Morrison. 1992. *Theory and Design of Adaptive Automation in Aviation Systems*. Technical Report ADA254595. Defense Technical Information Center. https://apps.dtic.mil/dtic/tr/fulltext/u2/a254595.pdf

[32] A. Rangesh, E. Ohn-Bar, and M.M. Trivedi. 2016. Long-Term Multi-Cue Tracking of Hands in Vehicles. *IEEE Transactions on Intelligent Transportation Systems* 17, 5 (May 2016), 1483–1492.

[33] A. Sathyanarayana, S. Nageswaren, H. Ghasemzadeh, R. Jafari, and J. H. L. Hansen. 2008. Body Sensor Networks for Driver Distraction Identification. In *2008 IEEE International Conference on Vehicular Electronics and Safety (VES*. 120–125.

[34] D. Saxby, G. Matthews, J. Warm, E. Hitchcock, and C. Neubauer. 2013. Active and Passive Fatigue in Simulated Driving: Discriminating Styles of Workload Regulation and their Safety Impacts. *Journal of Experimental Psychology* 19, 4 (2013), 287.

[35] D. Shepardson. 2019. Third Fatal Tesla Autopilot Crash Renews Questions About System. https://ca.reuters.com/article/technologyNews/idCAKCN1SM1QE-OCATC

[36] Y. Shiga, T. Toyama, Y. Utsumi, K. Kise, and A. Dengel. 2014. Daily Activity Recognition Combining Gaze Motion and Visual Features. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp): Adjunct Publication*. 1103–1111.

[37] I. Solís-Marcos, A. Galvao-Carmona, and K. Kircher. 2017. Reduced Attention Allocation during Short Periods of Partially Automated Driving: An Event-Related Potentials Study. *Frontiers in Human Neuroscience* 11 (2017), 537.

[38] R.M.A. van der Heiden, S.T. Iqbal, and C.P. Janssen. 2017. Priming Drivers Before Handover in Semi-Autonomous Cars. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI)*. 392–404.

[39] S. Vora, A. Rangesh, and M.M. Trivedi. 2018. Driver Gaze Zone Estimation Using Convolutional Neural Networks: A General Framework and Ablative Analysis. *IEEE Transactions on Intelligent Vehicles* 3, 3 (2018), 254–265.

[40] M. Walch, T. Sieber, P. Hock, M. Baumann, and M. Weber. 2016. 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)*. 261–268.

[41] M.S. Young and N.A. Stanton. 2002. Malleable Attentional Resources Theory: a New Explanation for the Effects of Mental Underload on Performance. *Human Factors* 44, 3 (2002), 365–375.