Is Too Much System Caution Counterproductive? Effects of Varying Sensitivity and Automation Levels in Vehicle Collision Avoidance Systems

Ernestine Fu
Stanford University
Stanford, California, USA
ernestinefu@stanford.edu

Mishel Johns
Stanford University
Stanford, California, USA
mishel@stanford.edu

David A. B. Hyde
UCLA
Los Angeles, California, USA
dabh@math.ucla.edu

Srinath Sibi
Stanford University
Stanford, California, USA
ssibi@stanford.edu

Martin Fischer
Stanford University
Stanford, California, USA
fischer@stanford.edu

David Sirkin
Stanford University
Stanford, California, USA
sirkin@stanford.edu

ABSTRACT

Autonomous vehicle system performance is limited by uncertainties inherent in the driving environment and challenges in processing sensor data. Engineers thus face the design decision of biasing systems toward lower sensitivity to potential threats (more misses) or higher sensitivity (more false alarms). We explored this problem for Automatic Emergency Braking systems in Level 3 autonomous vehicles, where the driver is required to monitor the system for failures. Participants (N=48) drove through a simulated suburban environment and experienced detection misses, perfect performance, or false alarms. We found that driver vigilance was greater for less-sensitive braking systems, resulting in improved performance during a potentially fatal failure. In addition, regardless of system bias, greater levels of autonomy resulted in significantly worse driver performance. Our results demonstrate that accounting for the effects of system bias on driver vigilance and performance will be critical design considerations as vehicle autonomy levels increase.

Author Keywords

Autonomous Vehicles; Automated Emergency Braking; Human Machine Interaction; Simulation; Controlled Experiment

CSS Concepts

• **Human-centered computing~Human computer interaction (HCI)**; User studies • **Interaction design**; Interface design prototyping

CHI '20, April 25–30, 2020, Honolulu, HI, USA

ACM 978-1-4503-6708-0/20/04 $15.00
https://doi.org/10.1145/3313831.3376300

Figure 1. Driving simulator scenario built to represent a Level 3 autonomous vehicle

INTRODUCTION

In a well-known illustration from the *Saturday Evening Post* from the 1950s, a glass bubble-topped car with large tailfins drives itself along a highway while its passengers relax over a game of dominoes [11]. While we are getting closer to realizing this prospect, autonomous vehicle systems in development today are still limited by their abilities to sense, process, interpret, and anticipate the full driving environment, and no system is entirely impervious to noise and misses. Any driving environment inherently contains a diversity of factors, such as other vehicles, pedestrians, and debris on the road—and the system needs to accurately detect and react to all of them. For instance, the system must properly classify a dog and avoid collision, while correctly rejecting a plastic bag blown by the wind [37]. Given these limitations, OEMs and engineers face the design challenge of biasing systems toward greater sensitivity and higher likelihood of false alarms (such as when the system triggers even though there is no collision threat), or toward lesser sensitivity and greater likelihood of a miss. The challenge becomes even more difficult as the vehicle system's level of autonomy increases, and consequently, driver vigilance and performance also change [20].

Automatic Emergency Braking (AEB) is one critical system component that designers must decide how to bias. AEBs can be present as either the single automated aspect of a vehicle or as one component of a higher-level autonomous system. In all levels of automation, when an obstacle and hazard is in the car's path and a collision is imminent, AEB can effectively provide an alert and automated braking, preventing collision independent of driver action [7, 37]. However, the AEB system can also pose a potential hazard, as when unnecessary braking occurs and causes a rear-end collision [55]. Moreover, when the AEB system provides too many false alarms, drivers can end up ignoring the AEB because it becomes a nuisance [3]. Complacency is yet another problem, whereby drivers become so dependent on the alert system that they rely completely on the system and fail to respond if the system fails [54].

This study explored the effects on drivers of various sensitivity levels of an AEB system present in Level 3 autonomous vehicles, where the system is mostly autonomous, but still requires some human operation and supervision. We investigated how various sensitivity settings for the AEB system influence driver awareness and performance. We build on the work of Fu et al., where participants were provided with a vehicle with a lower level of autonomy: a Level 1 autonomous vehicle, where the AEB was the only automated driver assist component [14]. Fu et al. found that drivers with an imperfect AEB were better at avoiding a critical collision when the AEB failed. To better understand how increases in autonomy affect driver performance, this study contained the same varying levels of system sensitivity as Fu et al., but used a vehicle capable of greater autonomy.

Participants were driven by a vehicle through a simulated suburban environment we programmed (see Figure 1), where they experienced a system that was biased either towards misses, perfect performance, or false alarms. Our findings suggest that driver vigilance is potentially greater with less sensitive AEB systems, whether the vehicle is a Level 1 or Level 3 system. We also observe that regardless of the AEB system's sensitivity level, higher levels of autonomy in vehicles result in reduced driver performance during potentially fatal events. These findings inform drivers and OEMs how to think about the implications of system bias, particularly as vehicles on the road with higher levels of automation become more prevalent. As the level of vehicle autonomy increases, but driver intervention is still required, we expect the effects of system bias on driver vigilance and performance to become more pronounced.

BACKGROUND

People form conceptual models and develop relationships of trust with interactive systems. These systems can range from service robots to autonomous vehicle systems [11, 47]. In general, the more a person interacts with a system, the stronger and more reliable that person's mental model is [20, 25, 33]. In the case of Level 3 autonomous vehicles with collision avoidance systems, the accuracy and trustworthiness of this mental model can have significant implications for human safety.

The simulation system with which users interacted in our study builds on the abilities of a vehicle with automated driving capabilities and an automated emergency braking system. The system also builds upon the literature of human behavior with varying levels of autonomous systems, signal detection theory, trust in and reliance on automation, as well as acquired complacency caused by alerting systems.

Human Vigilance and Performance with Varying Levels of Autonomous Systems

The human vigilance required for operating and overseeing an automated system varies with different levels of automation. The amount of oversight required of the human driver when it comes to intervention and attentiveness can affect such vigilance [20].

The idea of defining Levels of Automation (LOA) for systems traces to the seminal work of Sheridan and Verplank [38]. While some criticisms of the LOA framework have been made, academics and professionals are generally supportive of its practicality and utility for categorizing and designing autonomous systems [10, 23, 44, 45]. The Society of Automotive Engineers has defined five levels of automation, specifically for vehicles, which have been widely adopted by practitioners in the field [37]. These range from Level 1 systems, which provide driver assistance, up to Level 5 systems, which steer, accelerate, and monitor the environment fully autonomously without the need for human fallbacks. An AEB, for example, could be considered a complete Level 1 autonomous system or a component of a higher-level system [10].

An important design consideration of autonomous systems is how human vigilance changes as autonomy increases. Many of the worst accidents involving Level 2 and Level 3 autonomous vehicles currently deployed could have been prevented if the driver had been more attentive to the environment or driving task [16]. In fact, studies have shown that the driver of an autonomous vehicle can be as inattentive as a passenger in a human-driven vehicle, and that drivers can become physiologically and psychologically dependent on automation, resulting in less vigilance after only fifteen minutes of use of an autonomous system [1, 2, 48]. Although educating drivers about operating and interacting with autonomous systems may improve safety, system design should simultaneously be optimized [4]. However, studies have also identified findings such as the "prevalence paradox" of Sawyer and Hancock [39], which illustrates that when a system usually performs with high accuracy, a human user may be less vigilant in detecting and reacting to a system failure than if the system fails frequently. Additionally, heads-up display systems such as Google Glass, which aim to improve users' performance, may impair driver vigilance due to excessive information being provided [40].

While there are biometric and psychological means of assessing vigilance, a practical proxy for vigilance is measuring driver performance. For example, Sawyer and Hancock found that driving performance declined when users composed text messages through an automated assistive system, suggesting that the system impaired drivers' vigilance [38]. Johns et al. found that driving performance can vary significantly under varying levels of automation capabilities. When a transition from autonomous to manual control occurred, drivers were likely to suddenly apply much greater steering action when active steering or full autonomy was present, compared to only adaptive cruise control or full manual operation [20]. Johns et al. also concluded that low cognitive load can lead to vigilance decline and thereby impair driving performance. They even suggested that performing unrelated tasks, such as reading a book while piloting an autonomous vehicle, could increase cognitive load to the benefit of driving performance in the case of a transition of control to the driver [21].

Automated Emergency Braking Systems

Automated emergency braking (AEB) systems aim to prevent or significantly reduce the impact of frontal collisions. The potential life-saving benefits of AEB systems have led to their being required in passenger vehicles in the United States by 2022 [51]. However, these systems are challenging to design and implement; for instance, because of the AEB, rear-end collisions may occur with other vehicles, pedestrians, or other objects in roadways, and systems must be designed to handle such diverse obstacles while maintaining high accuracy. Despite advances in technologies such as RADAR, visual spectrum cameras, and sensor fusion used to power emergency braking systems, no emergency braking system is expected to have perfect accuracy [15, 26, 46]. Hence, there is an outstanding design question of whether to bias these systems towards reporting more erroneous false positives or towards being less sensitive to potential collision scenarios.

Since AEB systems are commonly restricted to detecting frontal collisions, there are clear risks associated with activating the system. In particular, the typically sudden nature of an emergency braking system activation can increase the likelihood of a rear-end collision if there is other traffic on the road; in some cases, this may be a worthwhile trade-off, but in the case of a false alarm, such braking clearly does more harm than if the system had not been activated [17]. On the other hand, under-reporting potential frontal collisions increases the risk of injury to the driver, pedestrians, and other drivers, and so a bias towards false alarms may be preferable to a bias towards underreporting [7]. Regardless, many AEB systems are designed to avoid a black-and-white decision of whether to activate; that is, a two-stage activation model is implemented whereby the system alerts, waits for driver action, and subsequently actuates the brakes only if the driver response is deemed insufficient [27]. This gives the driver an opportunity, albeit a brief one, to avoid an accident in a way that may be preferable to the course of action taken by the AEB system (such as steering away from danger or only partially actuating the brakes in heavy traffic). The simulated system used in this study follows this two-stage design of collision warning and automated braking.

Signal Detection Theory

Since our model AEB system may have false positives and/or false negatives, it is important to discuss signal detection theory (SDT), which studies the classification of and reaction to signals [30, 49]. A typical AEB system will predominantly detect noise (the lack of any impending collisions); however, it will occasionally detect a legitimate signal. SDT categorizes the presence or absence of a signal, along with whether the signal/noise is detected as a signal (see Table 1). In the case of autonomous driving systems, the signal detection matrix is repeated across two stages: first, the autonomous system will attempt to perform its duty, but if it fails, an overall successful outcome depends on the human operator performing any necessary driving maneuvers. Hence, both autonomous and human systems must fail to result in an overall signal detection failure, but the success of either system results in a signal being correctly detected.

	Signal Absent (noise)	**Signal Present (noise + signal)**
No Detection	Correct Rejection	Miss
Detection	False Alarm	Correct Detection (hit)

Table 1. Signal Detection Theory on the presence or absence of a signal and then the response to it

While ample research has investigated machine sensing and classification, as well as humans' abilities to detect and classify signals, we focused on the interplay of these two systems [50]. Our research was motivated by the question of how to bias an imperfect AEB system's evaluation of the four possible outcomes in order to increase overall success rates. We are interested in whether and how changes in the sensitivity of AEB and other autonomous systems; in other words, biases towards the different failure modes summarized in Table 1, affect drivers' behavior and particularly their reactions to potentially fatal accidents.

Automation, Trust and Reliance

Drivers need to trust autonomous systems to effectively use them. Neigel et al. demonstrated a correlation between trust in an autonomous system and task performance using the system [32]. Moreover, Mayer et al. characterize trust in automation as *"the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor"*; this definition includes risk of driver surprise from or driver disappointment in the imperfect performance of a system [31]. The risks focused on in this study are that a collision could occur if an AEB system fails to activate, and that accidents such as rear-end collisions could needlessly occur if the automated emergency braking system activates when there is no danger of a frontal collision.

When evaluating humans' trust in autonomous systems, it is useful to understand driver perception of such a system's reliability and trustworthiness, as well as their preconceived trust in technology at large. To measure these factors in our study, we used the questionnaire proposed by Jian et al. [18]. Users' mental models, and in particular trust, of a system can change through interaction with the system, and this trust can be measured by observing how much a user relies on a system over the course of repeated use.

Although sufficient trust is important, in the case of an AEB system, it is equally important that a user does not overtrust the system. Accidents due to overtrusting autonomous driving systems, which have been scrutinized in the media, highlight drivers' tendencies to maintain insufficient vigilance over the course of long drives, particularly in the presence of automation [29, 43]. Instead, users of AEB systems require calibrated trust based on an accurate perception of the system's performance. Lee and See clarify that this calibration is part of the user's mental model of the system and is developed via repeated interactions with the system [28]. In general, a user should ideally build a calibrated trust model based on many interactions with the system across a wide variety of scenarios, and the user should only rely on a system that has demonstrated a sufficiently acceptable level of calibrated trust. AEB systems present a particular challenge for this development of calibrated trust because an experienced driver will rarely encounter a positive signal from an AEB system. Hence, given that users will have minimal prior experience when such a system activates, the design and bias of automated emergency braking systems is especially important.

Alert Fatigue and Complacency

Alert fatigue occurs when users begin to ignore alerts from a system after too many false positives or non-actionable alerts have been issued. When a legitimate, actionable alert is produced by a system, users with alert fatigue are less likely to act upon the alert, diminishing the system's efficacy [9]. Similarly, if humans put too much trust in a system or are overwhelmed by many non-essential alerts, they can become complacent, which can lead to lack of vigilance, and fail to be ready to manually intervene in critical situations when systems fail to alert or critically alert. This complacent behavior may be characterized as a primary-secondary task inversion, where a user's primary task of paying attention is subsumed by the secondary task of passively monitoring for alerts and alarms [54]. Operators' alert fatigue and complacency have been thoroughly studied in the case of pilots and aviation [5, 6, 34].

Despite some similarities between piloting air and driving land vehicles, it is important to study alert fatigue and complacency within the automotive setting. Road and flight settings differ in significant ways: the degree of initial and ongoing training and experience for pilots versus drivers, the distances between nearby planes versus cars, and the time frames between alerts and collisions [42, 53]. For example, pilots may have several minutes to correct an issue with an aviation system if they are mid-flight; however, in the case of automobile accidents, drivers typically only have a few seconds at most to determine the best course of action in the face of a potentially fatal situation. These differences motivated us to consider that operator behavior and performance across contexts may vary significantly.

STUDY GOALS AND METHODS

We hypothesize that for a Level 3 autonomous system that fails to offer enough alerts when real hazards exist, drivers will increase their vigilance to compensate for the system's poor performance. Alternatively, if the system has perfect performance, drivers' vigilance will decrease, as complacency sets in, and thereby lower their abilities to respond to hazards that the system doesn't recognize or respond to. And if the system exhibits false alarms, we hypothesize that drivers' vigilance will decrease as they start to consider the system to be a distraction or nuisance, and thereby ignore it. We also hypothesize driver performance to worsen in Level 3 systems, given lower human vigilance required to operate the vehicle, in contrast to Level 1 systems.

To investigate our hypotheses, we designed a simulated driving experience that explored the formation and use of a metal model of system performance by biasing towards different failure modes, and we compared how such varying sensitivity levels in systems affects driver performance in a potentially fatal failure. We then compared our results with those of Fu et al., which studied AEB sensitivity levels in a lower-level automated system [14].

Participants

Participants between the age of 18 to 60 years old (*M* = 26.72 years, *Mdn* = 23 years, *SD* = 8.41 years) were recruited using flyers and emails and were compensated for their time with an Amazon gift card. Participants' driving experience ranged from 2 to 43 years (*M* = 10.71 years, *Mdn* = 8 years, *SD* = 8.95 years). Participants reported driving between 0.3 and 7 days per week (*M* = 4.02 days, *Mdn* = 4 days, *SD* = 2.59 days).

	Misses	**Perfect**	**Perfect**	**False Alarms**
Training Course	3 misses 6 correct detections	9 correct detections	9 correct detections	3 false alarms 6 correct detections
Final Event	Detection Failure		Brake Failure	
Number of Participants	N = 12	N = 12	N = 12	N = 12

Table 2. Experimental conditions

Driving Simulator and Study Context

The study was conducted in a fixed-base driving simulator using a full-vehicle cab, 270° wrap-around screen, rear-view screen, separate video channels for rear view mirrors as well as full audio system (see Figure 1).

Participants were provided with a Level 3 autonomous vehicle that contained an emergency braking system with varying levels of sensitivity. The experiment comprised two sections: a training course and a final event (see Table 2). In the training course, the vehicle encountered nine pedestrian events—some were potential hazards, while others were not. Participants were assigned to one of three training groups, which dictated how the vehicle responded to the pedestrians: under-sensitive system biased towards misses, over-sensitive system biased towards false alarms, or perfect performance. This design allowed participants to form a mental model of the system's capabilities and sensitivity level. The final event was split into two conditions: participants experienced either a detection failure, where the car failed to provide an automatic alert and brake, or a brake system failure, where the car provided an alert, but did not apply automatic braking. This design allowed for a pairwise comparison of perfect performance with misses, and then perfect performance with false alarms.

	Misses with Detection Failure	**Perfect with Detection Failure**	**Perfect with Brake Failure**	**False Alarms with Brake Failure**
Correct Detections	3	6	6	6
Misses	3	0	0	0
False Alarms	0	0	0	3
Correct Rejections	3	3	3	0

Table 3. The number of simulation events, and system performance in the training course by condition (as columns)

Course and Procedure

Participants were presented with a course that took approximately 45 minutes to complete. The course contained segments where participants had to drive manually, as well as segments where the vehicle's automated driving system took control (see Figure 2).

So that they could familiarize themselves with the simulated driving environment and how to operate the vehicle, participants first drove for approximately 4 minutes in a course section that contained an assortment of road types, as well as audio instructions to enable and disable automated driving.

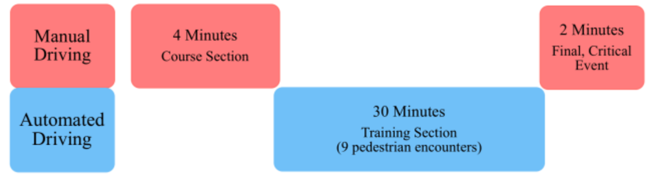

Figure 2. Diagram of the simulated driving course

At the end of this initial course section, participants were asked to enable automated driving as they entered a training section that consisted of approximately 30 minutes of driving. The training section consisted of nine pedestrian incursion events. We asked participants to allow for automation to perform the majority of the driving but told them that they could still take control of the car at any time if they felt in danger. Participants could disengage automation during the drive by either stepping on the brakes or turning the steering wheel at least 15 degrees. In addition to driving autonomously, the simulated vehicle included an automated forward collision warning system and automated emergency braking. The system provided a verbal alert "Warning: Obstacle Detected" and initiated an automated braking action when detecting a hazard.

During the training section, pedestrians crossed the street directly in the vehicle's path six times: from the right to left-hand side of the road four times, and from the left to right-hand side two times. Pedestrians also ran along the sidewalk, not impeding the vehicle's path, three times: on the right-hand side of the road two times, and on the left-hand side once (see Table 3).

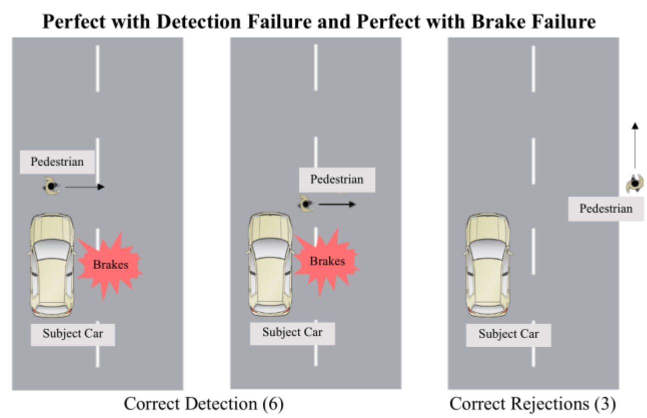

Figure 3. System exhibiting perfect performance with either detection failure or brake failure in the final event

When the system exhibited perfect performance, it always worked accurately. The car notified and braked during all six times a pedestrian crossed the street, and when a pedestrian was simply walking on the sidewalk, the system correctly did not take action (see Figure 3).

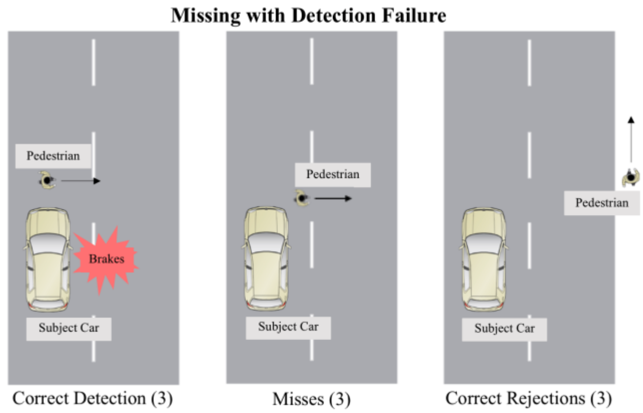

Figure 4. System exhibiting misses with detection failure in the final event

When the system exhibited misses, it failed to detect some pedestrians. Specifically, there were three instances where the car did not alert and brake during a pedestrian crossing. To reflect existing industry design, there were also some instances where the car exhibited correct performance (see Figure 4).

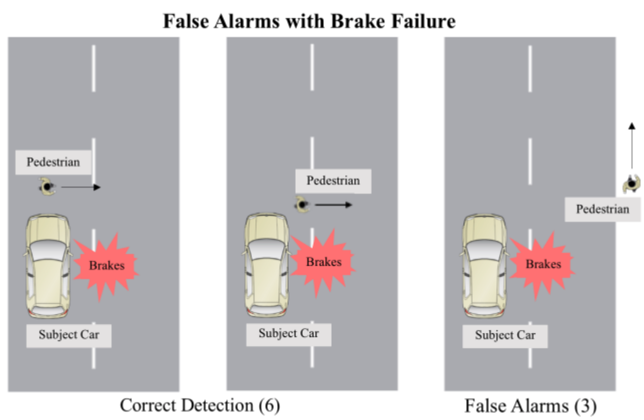

Figure 5. System exhibiting false alarms with brake failure in the final event

When the system exhibited false alarms, it issued an alert and applied the brakes—even if a pedestrian was not a threat. There were three instances of a pedestrian simply walking along the sidewalk and not crossing the vehicle's path, but the system nevertheless provided an alert and braking. When a pedestrian did cross the road, the system also always took action (see Figure 5).

Figure 6. Final event requires participants to disengage automation to navigate a pedestrian and dog crossing the road, as the automation system fails

During the final event section, all participants were presented with a tenth pedestrian encounter. The pedestrian hazard in this event included a person walking with a dog across the street (see Figure 6). This event was used as the main metric to measure changes in driver performance between conditions.

There were two types of failures presented to participants. Some participants experienced a detection failure, in which the car neither braked nor provided an alert. Other participants experienced a brake system failure, where the car failed to brake even though it correctly provided an automated alert. Together, these distinct failure types allowed us to analyze the four different conditions experienced by participants.

RESULTS

We focus our analysis on the final event to understand participant behavior and reliance on the system during a critical failure. Performance in avoiding a collision, vehicle speed, and reaction time varied based on the sensitivity level of the AEB system that the driver was provided. The following analyses compare the following condition pairs: Missing versus Perfect with Detection Failure, Perfect with Detection Failure versus Perfect with Brake Failure, and Perfect with Brake Failure versus False Alarms.

We then compare results from our study's Level 3 autonomous vehicle with that of Fu et al.'s results for a Level 1 autonomous vehicle, where participants manually drove themselves (instead of being driven by the autonomous system), encountered the same set of obstacles, and experienced an AEB system with varying sensitivity levels [14]. The comparison focuses on both participant performance and vehicle speed between the Level 1 and Level 3 system for the final, critical event.

Participant Performance in Critical Event with Level 3 System

The number of participants who did not collide with the pedestrian walking a dog was highest in the Missing condition (10). In the other three conditions, an equal number of participants navigated the final event successfully (3) (see Figure 7).

We used Fisher's exact test to analyze success and failure for the final event under the four conditions. For the small fraction of users who navigated the final critical event successfully, this test yielded a statistically significant difference between the Missing condition (10/12) and Perfect with Detection Failure condition (3/12) ($p = 0.01$).

We note that the Missing condition would also show significant differences if it were compared to the other Perfect with Brake Failure and False Alarms conditions, but the conditions are not directly comparable, as doing so would result in two changing variables.

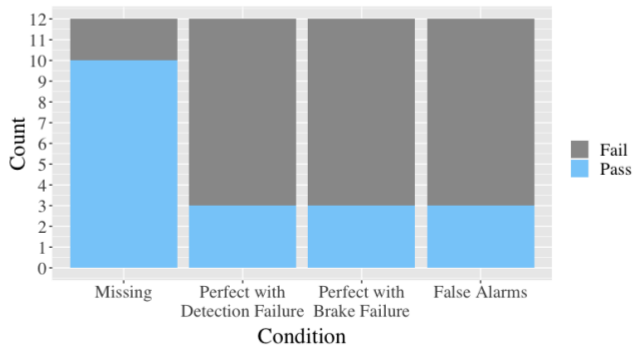

Figure 7. Participant performance of failure or success in navigating the final event

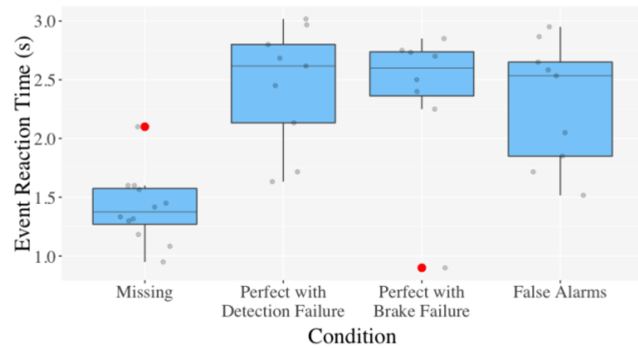

Figure 9. Participant's reaction time to disengage automation in the final event

Vehicle Speed in Critical Event with Level 3 System

We analyzed the vehicle's speed as it passed by or collided with the pedestrian walking a dog (see Figure 8). If the participant failed to disengage automation, the car collided into the pedestrian and dog at its automated driving speed of 20 m/s. This speed value thus served as a proxy for whether a participant disengaged automation and slowed to avoid the pedestrian or not. We evaluated this metric across each of the four conditions. We then ran an independent-samples t-test across the four data sets, which yielded statistically significant differences between the Missing (*M*=3.56 m/s, *SD*=4.37 m/s) and Perfect with Detection Failure (*M*=16.88 m/s, *SD*=3.2 m/s) conditions; $t(23)=-8.529$, $p < 0.001$. The results suggest that participants in the Missing condition successfully slowed the vehicle down to avoid the impending collision.

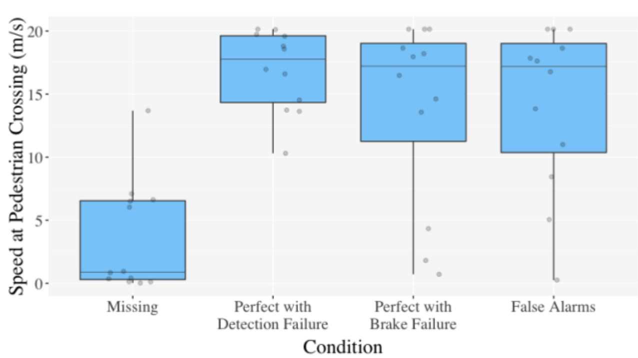

Figure 8. Vehicle's speed as vehicle encounters the pedestrian and dog

Reaction Time in Critical Event with Level 3 System

We analyzed the reaction time to understand when participants either engaged the vehicle's brakes or adjusted the steering wheel at least 15 degrees in order to disengage automation before the final, critical event (see Figure 9). We calculated reaction time from the moment the pedestrian and dog started moving across the street.

Using pairwise t-tests, we found a significant difference in the reaction time for Missing (*M*=1.41 s, *SD*=0.30 s) and Perfect with Detection Failure (*M*=2.45 s, *SD*=0.51 s) conditions; $t(23)=-5.4338$, $p < 0.001$. These results suggest that participants in the Missing condition disengaged automation in a shorter period of time in the final event.

Trust Measure in Level 3 System

To evaluate trust between the participant and automation system, we examined their behavioral demonstration of reliance or non-reliance on the system's behaviors using Jian et al.'s self-reported measures questionnaire [18], an empirically based scale used by researchers studying trust in autonomous vehicles. Before and after the simulated driving experience, participants completed the questionnaire to rate their feelings of trust and impressions of the automation system. A repeated-measures ANOVA did not find statistically significant difference between the pre- and post-drive trust index, nor was there a significant change in trust scores among the four experimental conditions (see Figure 10).

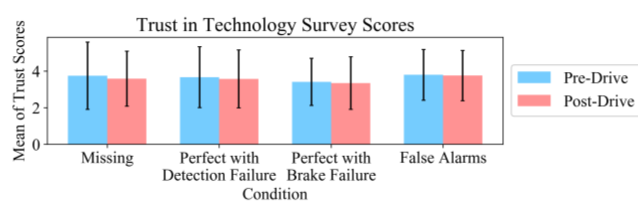

Figure 10. Self-reported measurement of trust in the automation system

Participant Performance Comparison of Level 1 (Manual-Configuration) and Level 3 (Automation-Configuration) Systems

The metrics discussed in the previous section were compared, in the same pairs, to the results from data we obtained from Fu et al., where participants were instructed to manually drive through the same obstacle and road path using a Level 1 vehicle, where the AEB system was the only advanced driver assist system [14]. Given the similar condition groups used to categorize participants in both studies, we compared the following pairs in both the Automation and Manual studies: Missing versus Perfect with Detection Failure, Perfect with Detection Failure versus

Perfect with Brake Failure, and Perfect with Brake Failure versus False Alarms.

Overall, far fewer participants successfully navigated the final event in the Automation condition (19) compared to the Manual condition (35) (see Figure 11). In particular, performance was significantly worse in the Perfect with Brake Failure and False Alarm conditions, but not in the Perfect with Detection Failure and Missing conditions.

To quantify the difference between the results of the current and previous study, we computed Fisher's exact test statistics for each of the four pairs of results across the two studies, one pair per condition. The Perfect with Brake Failure condition results between the Automation configuration (3/12) and Manual configuration (11/12) gave a statistic of 33.0 (p = .003), indicating a significant difference between the two studies. The False Alarm condition between the Automation configuration (3/12) and Manual configuration (10/12) also showed significant differences between the two studies, with a statistic of 15.0 (p = .012).

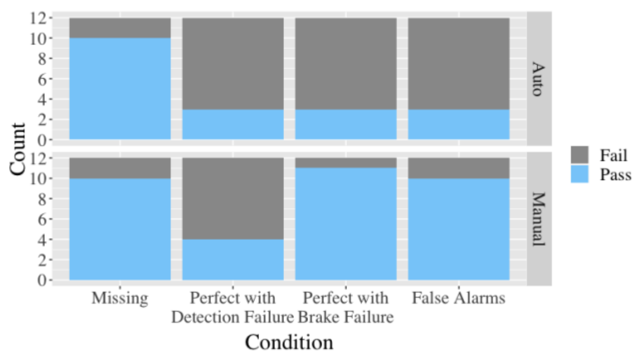

Figure 11. Comparison of participant performance for the manual and automation configurations

Vehicle Speed Comparison of Level 1 and Level 3 Systems

We performed a similar analysis to compare vehicle speed at collision time across the Automation and Manual configurations, for each of the four conditions. Since vehicle speeds are continuous variables rather than discrete (binary) events, we used the two-sample Kolmogorov-Smirnov test [8], which answers whether two continuous data sets are likely to have been drawn from the same distribution.

For the Perfect with Detection Failure condition, a statistic of 0.67 was obtained (p = .008), which indicates a significant difference for the vehicle speeds between the Automated and Manual configurations. We can conclude that there is significantly different behavior in the Perfect with Detection Failure condition between participants driving the vehicle themselves and the car autonomously driving participants. For the Perfect with Brake Failure and False Alarm conditions, identical results were obtained when comparing the Automated and Manual configurations, with a statistic value of 0.75 (p = .002); both conditions displayed significantly different results across the two studies' automation results. Overall, only the Missing condition demonstrated consistently insignificant results, suggesting that the Automated configuration results in significantly different behavior when the ADAS provides perfect or excessive false signals to the user.

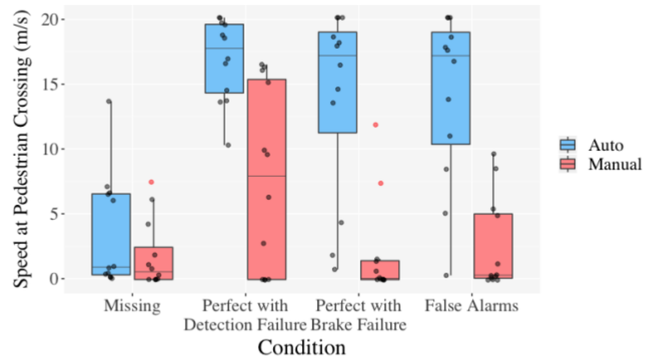

Figure 12. Comparison of vehicle speed during final event for the manual and automation configurations

DISCUSSION AND CONCLUSIONS

The driving performance of participants, along with their reaction time to disengage automation and thereby the vehicle's speed, varied when they were presented with a Level 3 autonomous vehicle's AEB system that signaled either fewer or more false alarms and misses.

In the Missing condition, most participants successfully navigated the final event. They tended to disengage automation and engage the vehicle's brakes early enough so that when the pedestrian and dog crossed the road, the vehicle was at a slower speed, occasionally even at 0 m/s, a full stop. Their reaction times also tended to be lower, which means they were able to disengage automation in a shorter time period when the vehicle encountered a threat in the final event. We expect that participants were more cautious when using a vehicle that already exhibited some errors and misses during the training course (see Figure 13), and therefore were more alert and cautious during the final event. In the video data, Missing condition participants were observed to be frequently leaning in and checking their surroundings while on the road.

Figure 13. Front and aerial view of less sensitive system that does not detect and brake for pedestrians in opposing traffic lane

In the two Perfect conditions and also False Alarms condition, most participants were unable to navigate the final event. Many collisions occurred at relatively high speeds, approaching 20 m/s, which meant that the driver did not bother to disengage automation and the vehicle maintained its existing automated speed when colliding into the final event's pedestrian and dog. This behavior was further confirmed by the reaction time metric, where the driver reacted several seconds after the pedestrian and dog started crossing the street, sometimes disengaging automation only after the fatal collision occurred. We expect that this behavior resulted from a sense of passivity and complacency in the vehicle's performance, as prior to the final, critical event, the automated system successfully navigated any obstacle that was in its path. In the False Alarms condition, the vehicle even slowed down, alerted and braked for objects that were on sidewalks, even though they were not in the vehicle's path (see Figure 14). Participants in the Perfect conditions and False Alarms condition may have had a sense of over-confidence in the vehicle's ability to accurately detect any and all potential threats—even if they were not obstacles on the direct road path—and therefore, when the vehicle was unable to correctly react and perform in the final event, participants were not able to successfully take over.

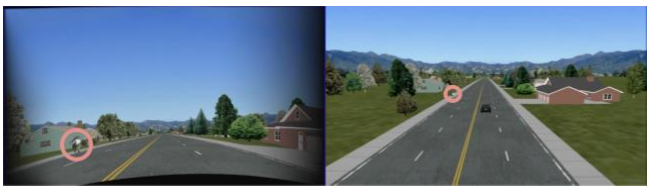

Figure 14. Front and aerial view of system biased towards false alarms brakes for all potential threats, including pedestrians on sidewalk who are not in the direct road path

When comparing the Manual and Automated configurations, we noticed shifts in participant behavior as the vehicle's level of automation increased. Overall, far fewer participants successfully navigated the final event in the Automated configuration compared to the Manual configuration. In the exact same condition groups of varying levels of AEB sensitivity, we noticed better performance and higher vigilance when participants were in the Manual configuration. The most significant difference occurred in the Perfect with Brake Failure and False Alarms conditions, where the Automated configuration resulted in significantly worse performance. We expect that this difference occurred because there is naturally a higher level of vigilance when one is driving a vehicle compared to when one is simply supervising. Drivers who merely supervise the vehicle may come to view themselves more as passengers, without needing to take any action, especially after they have formed a mental model that the vehicle can handle both real and potential threats by itself. And finally, whether the participant received the Manual or Automated configuration, participants in the Missing condition resulted in better driver performance. We expect that this behavior occurred because when a system is biased towards a failure mode that may involve fatal accidents—in other words, misses—drivers become extremely wary and cautious of the system, even more so than when the car is biased towards false alarms. The implication is that a system biased towards misses increases driver vigilance and improves driver performance in cases of detection failure or system inaction.

Our findings suggest a challenge for automated vehicle system designers: that an AEB system biased toward issuing fewer alerts, even when the threats are real (an undesirable design specification), results in increased driver vigilance and performance (a desirable outcome). We expect that as autonomous vehicle sensing and control systems become capable of handling all situations without driver supervision required, as in Level 5 systems, this challenge will recede, because vehicle occupants will not need to maintain high degrees of situation awareness. However, in the interim, as the level of vehicle autonomy increases, yet the need for drivers to potentially intervene remains, the effects of system bias on driver vigilance and performance may become more pronounced.

In conclusion, we had hypothesized that a system exhibiting misses can result in improved driver response during critical events, and that a system with perfect performance or false alarms would lead to complacency that negatively influenced driver response. Our study supports the first hypothesis, as drivers reacted more to a system biased to under-report hazards. We also hypothesized that higher levels of autonomy in vehicles result in a lower level of driver vigilance and awareness. When reversing the roles of the driver and computer and tasking the driver to supervise an imperfect higher-level automated system, we noticed that driver performance worsened during a final, critical event.

DESIGN IMPLICATIONS

The design of systems with misses and false alarms has been applied in other domains, e.g., inventory control (rejecting high-quality goods and accepting low-quality goods) [24], computer security (classifying imposters as authorized users and authorized users as imposters) [41], airport security screening (identifying an innocent traveler as a terrorist) [13], and biometrics and medical testing [16]. As with autonomous vehicles, these systems have inherent algorithmic biases, and the consequences of biasing towards more misses or more false alarms may have more severe—sometimes dire—consequences, depending on the application; for example, indicating a woman is not pregnant when she is, or acquitting a guilty person of a crime.

Our findings provide insights to car manufacturers regarding the design of automotive system bias. The default path that companies may be inclined to take is conservative; however, the interpretation of 'conservative' may differ between manufacturers. One manufacturer may provide an alert whenever a situation might warrant it (as the threat could be real), while another may not provide one alert after another unless the system is absolutely certain the threat is real. Both are cautious frames, yet both result in different implementations. The resulting inconsistency across vehicles can cause confusion among drivers, as they move from one manufacturer's vehicles to another's, and our findings provide a starting point to address such design challenges.

Still, we hesitate to suggest that our findings be immediately engineered into active systems. Rather, we encourage additional design research to assess if there are other means to train a driver to expect misses and under-notification in autonomous vehicles. This may occur through new driving tests for operating autonomous vehicles, or through training of the driver that system flaws will occur.

LIMITATIONS AND FUTURE WORK

One shortcoming of simulator studies is that driver trust in autonomous systems is likely different compared to that in autonomous vehicles on the road with real traffic. However, given the dangers of conducting studies with critical safety failure events on the open road, simulator-based studies are an ethical compromise. We hope to modify the study design in the future to allow experiments on the physical road.

Additionally, our study is based on standard signal detection theory, which has received the greatest attention in detection theory literature; however, there are related theories worth noting. In particular, fuzzy SDT has been applied to studying both human and machine performance, and in particular, it has been used to evaluate models for human drivers' ability to perceive hazards [35, 52]. Fuzzy SDT relaxes the black-and-white categorizations of signal versus noise and detection versus miss and can more robustly model partial failures, such as when a driver brakes enough to avoid a fatality but not enough to avoid a collision. While we focus on standard SDT, we hope to expand the study design in the future to include fuzzy SDT concepts.

While we did not observe changes in participants' trust in automation using pre- and post-study questionnaires, we expect that additional probes of trust during the drive could reveal significant differences in future studies. Additionally, while we focused on driver performance as a main metric and observed significant differences, we also encourage research on the psychological and economic outcomes of designing systems with more misses, as has been done with research on designing biases in cancer detection screening [36].

In sum, while there has been a great deal of research in the area of automation control, further study is required to offer guidance to drivers and OEMs about how to think about the implications of system bias, particularly as vehicles on the road with higher levels of automation—such as the car depicted in the 1950s illustration of future highway travel—become more prevalent.

ACKNOWLEDGEMENTS

This study was conducted under Stanford University IRB Protocol 30016. We thank our fellow researchers at Stanford's Center for Design Research for their advice in the development of the study.

REFERENCES

[1] Arakawa, Toshiya. 2017. Trial verification of human reliance on autonomous vehicles from the viewpoint of human factors. *International Journal of Innovative Computing, Information and Control* 14, 2 (April 2018), 491–501. http://www.ijicic.org/ijicic-140207.pdf

[2] Arakawa, Toshiya and Kunihiko Oi. (2016). Verification of autonomous vehicle over-reliance. In *Proceedings of the Measuring Behavior* 2016, 177–182. https://www.measuringbehavior.org/mb2016/files/2016/MB2016_Proceedings.pdf

[3] Breznitz, Shlomo. 1984. *Cry wolf: the psychology of false alarms*. Lawrence Erlbaum Associates.

[4] Casner, Stephen M. and Edwin L. Hutchins. 2019. What do we tell the drivers? Toward minimum driver training standards for partially automated cars. *Journal of Cognitive Engineering and Decision Making* 13, 2 (2019), 55–66. http://doi.org/10.1177/1555343419830901

[5] Casner, Stephen M. and Jonathan W. Schooler. 2013. Thoughts in flight: automation use and pilots' task-related and task-unrelated thought. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 56, 3 (2013), 433–442. http://doi.org/10.1177/0018720813501550

[6] Casner, Stephen M., Richard W. Geven, Matthias P. Recker, and Jonathan W. Schooler. 2014. The retention of manual flying skills in the automated cockpit. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 56, 8 (2014), 1506–1516. http://doi.org/10.1177/0018720814535628

[7] Coelingh, Erik, Andreas Eidehall, and Mattias Bengtsson. 2010. Collision Warning with Full Auto Brake and Pedestrian Detection - A practical Example of Automatic Emergency Braking. *13th International IEEE Conference on Intelligent Transportation Systems* (2010), 155-160. http://doi.org/10.1109/itsc.2010.5625077

[8] Corder, Gregory W. and Dale I. Foreman. 2014. *Nonparametric Statistics: A Step-by-Step Approach*. John Wiley & Sons.

[9] Cummings, M.L., Ryan M. Kilgore, Enlie Wang, Louis Tijerina, and Dev S. Kochhar. 2007. Effects of single versus multiple warnings on driver performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49, 6 (2007), 1097–1106. http://doi.org/10.1518/001872007x249956

[10] Endsley, Mica R. 2017. Level of automation forms a key aspect of autonomy design. *Journal of Cognitive Engineering and Decision Making* 12, 1 (2018), 29–34. http://doi.org/10.1177/1555343417723432

[11] Everett Collection. 1950s. Driverless Car of the Future, America's Electric Light and Power Companies, *Saturday Evening Post*. Retrieved August 16, 2019 from https://www.computerhistory.org/atchm/where-to-a-history-of-autonomous-vehicles/

[12] Forlizzi, Jodi and Carl Disalvo. 2006. Service robots in the domestic environment: A study of the roomba vacuum in the home. In *Proceeding of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction - HRI 06*. http://doi.org/10.1145/1121241.1121286

[13] Frederickson, H. George and Todd R. Laporte. 2002. Airport security, high reliability, and the problem of rationality. *Public Administration Review* 62, s1 (2002), 33–43. http://doi.org/10.1111/1540-6210.62.s1.7

[14] Fu, Ernestine, Srinath Sibi, David Miller, Mishel Johns, Brian Mok, Martin Fischer, and David Sirkin. 2019. The car that cried wolf: Driver responses to missing, perfectly performing, and oversensitive collision avoidance systems. *2019 IEEE Intelligent Vehicles Symposium (IV)*, 1830-1836. http://doi.org/10.1109/ivs.2019.8814190

[15] Gade, Rikke and Thomas B. Moeslund. 2013. Thermal cameras and applications: A survey. *Machine Vision and Applications* 25, 1 (2013), 245–262. http://doi.org/10.1007/s00138-013-0570-5

[16] Hall, Sue, Martin Bobrow, and Theresa M. Marteau. 2000. Psychological consequences for parents of false negative results on prenatal screening for Downs syndrome: Retrospective interview study. *BMJ* 320 (2000), 407–412. http://doi.org/10.1136/bmj.320.7232.407

[17] Hubele, Norma and Kathryn Kennedy. 2018. Forward collision warning system impact. *Traffic Injury Prevention* 19, sup2. http://doi.org/10.1080/15389588.2018.1490020

[18] Jian, Jiun-Yin, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. http://doi.org/10.1207/s15327566ijce0401_04

[19] Jiménez, Felipe, José Naranjo, Sofía Sánchez, et al. 2018. Communications and driver monitoring aids for fostering SAE level-4 road vehicles automation. *Electronics* 7, 10 (2018), 228. http://doi.org/10.3390/electronics7100228

[20] Johns, Mishel, David B Miller, Annabel C Sun, Shawnee Baughman, Tongda Zhang, and Wendy Ju. 2015. The driver has control: Exploring driving performance with varying automation capabilities. In *Proceedings of the 8th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design: driving assessment 2015*. http://doi.org/10.17077/drivingassessment.1600

[21] Johns, Mishel, Srinath Sibi, and Wendy Ju. 2014. Effect of cognitive load in autonomous vehicles on driver performance during transfer of control. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI 14*. http://doi.org/10.1145/2667239.2667296

[22] Johnson-Laird, P.N. 1980. Mental models in cognitive science. *Cognitive Science* 4, 1 (1980), 71–115. http://doi.org/10.1207/s15516709cog0401_4

[23] Kaber, David B. 2017. Issues in human - automation interaction modeling: Presumptive aspects of frameworks of types and levels of automation. Journal of Cognitive Engineering and Decision Making 12, 1 (2018), 7–24. http://doi.org/10.1177/1555343417737203

[24] Kawanaka, Tenma and Tsukasa Kudo. 2018. Inventory satisfaction discrimination method utilizing images and deep learning. *Procedia Computer Science* 126 (2018), 937–946. http://doi.org/10.1016/j.procs.2018.08.028

[25] Kieras, David E. and Susan Bovair. 1984. The role of a mental model in learning to operate a device. *Cognitive Science* 8, 3 (1984), 255–273. http://doi.org/10.1207/s15516709cog0803_3

[26] Kivelevitch, Elad H., Greg Dionne, Trevor Roose, et al. 2019. Sensor Fusion Tools in Support of Autonomous Systems. *AIAA Scitech 2019 Forum*. http://doi.org/10.2514/6.2019-0384

[27] Lee, Hyuck-Kee, Seong-Geun Shin, and Dong-Soo Kwon. 2017. Design of emergency braking algorithm for pedestrian protection based on multi-sensor fusion. *International Journal of Automotive Technology* 18, 6 (2017), 1067–1076. http://doi.org/10.1007/s12239-017-0104-7

[28] Lee, John D. and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (2004), 50–80. http://doi.org/10.1518/hfes.46.1.50.30392

[29] Lin, Patrick. 2016. Tesla Autopilot Crash: Why We Should Worry about a Single Death. *IEEE Spectrum: Technology, Engineering, and Science News*. 2016. Retrieved September 8, 2019 from https://spectrum.ieee.org/cars-that-think/transportation/self-driving/tesla-autopilot-crash-why-we-should-worry-about-a-single-death

[30] Macmillan, Neil A. 2002. Signal detection theory. In Stevens' *Handbook of Experimental Psychology, Vol. 4, Methodology in Experimental Psychology* (3rd. ed.), Hal Pashler and John T. Wixted (eds.). John Wiley & Sons, NY, 42-90.

[31] Mayer, Roger C., James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *The Academy of Management Review* 20, 3 (1995), 709. http://doi.org/10.2307/258792

[32] Neigel, Alexis R., Justine P. Caylor, Sue E. Kase, Michelle T. Vanni, and Jefferson Hoye. 2018. The role of trust and automation in an intelligence analyst decisional guidance paradigm. *Journal of Cognitive Engineering and Decision Making* 12, 4 (2018), 239–247. http://doi.org/10.1177/1555343418799601

[33] Norman, Donald A. 1982. Some observations on mental models. In *Mental Models*, Dedre Gentner and Albert L. Stevens (eds.). Lawrence Erlbaum Associates, Hillsdale, NJ, 7-14.

[34] Palmer, Everett A., Edwin L. Hutchins, Richard D. Ritter, and Inge vanCleemput. (1993). Altitude Deviations: Breakdowns of an Error-Tolerant System. NASA Technical Memorandum 108788. http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940011077.pdf

[35] Parasuraman, Raja, Anthony J. Masalonis, and Peter A. Hancock. 2000. Fuzzy signal detection theory: Basic postulates and formulas for analyzing human and machine performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 42, 4 (2000), 636–659. http://doi.org/10.1518/001872000779697980

[36] Petticrew, M., A. Sowden, D. Lister-Sharp, and K. Wright. 2000. False-negative results in screening programmes: Systematic review of impact and implications. *Health Technology Assessment* 4, 5 (2000), 1-120. http://doi.org/10.3310/hta4050

[37] SAE. 2014. Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems. SAE Standard J3016_201401. http://doi.org/10.4271/j3016_201401

[38] Sawyer, Ben D. and Peter A Hancock. 2013. Performance degradation due to automation in texting while driving. In *Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design: driving assessment 2013*. http://doi.org/10.17077/drivingassessment.1525

[39] Sawyer, Ben D. and Peter A. Hancock. 2018. Hacking the human: The prevalence paradox in cybersecurity. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 60, 5 (2018), 597–609. http://doi.org/10.1177/0018720818780472

[40] Sawyer, Ben D., Victor S. Finomore, Andres A. Calvo, and P. A. Hancock. 2014. Google Glass: A driver distraction cause or cure? *Human Factors: The Journal of the Human Factors and Ergonomics Society* 56, 7 (2014), 1307–1321. http://doi.org/10.1177/0018720814555723

[41] Shanmugapriya, D., and G. Padmavathi. 2009. A Survey of Biometric Keystroke Dynamics: Approaches, Security and Challenges. https://arxiv.org/ftp/arxiv/papers/0910/0910.0817.pdf

[42] Sheridan, Thomas B. 2002. *Humans and Automation: System Design and Research Issues*. John Wiley & Sons, NY.

[43] Sheridan, Thomas B. 2006. Supervisory control. In *Handbook of Human Factors and Ergonomics* (3rd. ed.), Gavriel Salvendy (ed.). John Wiley & Sons, Hoboken, NJ, 1025-052. https://doi.org/10.1002/0470048204.ch38

[44] Sheridan, Thomas B. 2017. Comments on “Issues in human–automation interaction modeling: Presumptive aspects of frameworks of types and levels of automation” by David B. Kaber. *Journal of Cognitive Engineering and Decision Making* 12, 1 (2018), 25–28. http://doi.org/10.1177/1555343417724964

[45] Sheridan, Thomas B. and William L. Verplank. 1978. Human and Computer Control of Undersea Teleoperators. Technical Report. http://doi.org/10.21236/ada057655

[46] Singh, Additi Mrinal, Soumyasree Bera, and Rabindranath Bera. 2018. Review on Vehicular Radar for Road Safety. *Advances in Communication, Cloud, and Big Data Lecture Notes in Networks and Systems*, 41- 47. http://doi.org/10.1007/978-981-10-8911-4_5

[47] Sung, Ja-Young, Lan Guo, Rebecca E. Grinter, and Henrik I. Christensen. (2007). “My Roomba is Rambo”: Intimate home appliances. *UbiComp 2007: Ubiquitous Computing Lecture Notes in Computer Science*, 145–162. http://doi.org/10.1007/978-3-540-74853-3_9

[48] Takeda, Yuji, Toshihisa Sato, Kenta Kimura, Hidehiko Komine, Motoyuki Akamatsu, and Jun Sato. 2016. Electrophysiological evaluation of attention in drivers and passengers: Toward an understanding of drivers’ attentional state in autonomous vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour* 42 (2016), 140–150. http://doi.org/10.1016/j.trf.2016.07.008

[49] Tanner, Wilson P. and John A. Swets. 1954. A decision-making theory of visual detection. *Psychological Review* 61, 6 (1954), 401–409. http://doi.org/10.1037/h0058700

[50] Tuzlukov, Vyacheslav P. 2013. *Signal Detection Theory*. Springer Science & Business Media.

[51] U.S. DOT and IIHS Announce Historic Commitment of 20 Automakers to Make Automatic Emergency Braking Standard on New Vehicles. 2016. *IIHS*. Retrieved September 8, 2019 from https://www.iihs.org/news/detail/u-s-dot-and-iihs-announce-historic-commitment-of-20-automakers-to-make-automatic-emergency-braking-standard-on-new-vehicles

[52] Wallis, Thomas S.A. and Mark S. Horswill. 2007. Using fuzzy signal detection theory to determine why

experienced and trained drivers respond faster than novices in a hazard perception test. *Accident Analysis & Prevention* 39, 6 (2007), 1177–1185. http://doi.org/10.1016/j.aap.2007.03.003

[53] Wheeler, W. A. and T. J. Trigs. 1995. A task analytical view of simulator based training for drivers. *Road Safety Research and Enforcement Conference* (1996), Coogee Beach, New South Wales, Australia.

[54] Wiener, Earl L. 1985. Human Factors of Cockpit Automation: A Field Study of Flight Crew Transition. NASA Contractor Report 177333. http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19850021625.pdf

[55] Xia, Likun, Tran Duc Chung, and Khairil Anwar Bin Abu Kassim. 2013. A review of automated emergency braking system and the trending for future vehicles. In *Proceedings of the Southeast Asia Safer Mobility Symposium 2013* (SAEM 2013-010). https://www.saemalaysia.org.my/wp-content/uploads/2017/03/SAEM-2013-010.pdf