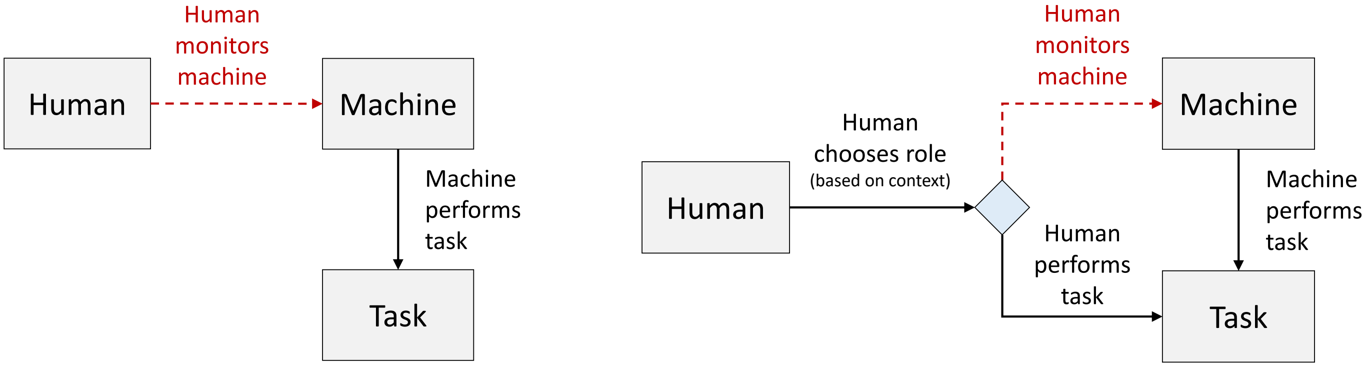
Human Side of Tesla Autopilot: Exploration of Functional Vigilance in Real-World Human-Machine Collaboration

Lex Fridman*[a] Daniel E. Brown[a] Julia Kindelsberger[a]
Linda Angell[b] Bruce Mehler[a] Bryan Reimer[a]

[a] Center for Transportation and Logistics,
Massachusetts Institute of Technology (MIT)
[b] Touchstone Evaluations, Inc.

***Abstract*—Over seventy years of research into human behavior in the context of automation shows that humans naturally over-trust reliable automation after relinquishing control for prolonged periods of time. This paper aims to show, by using an analysis of real-world driving in Autopilot-equipped Tesla vehicles, that patterns of decreased vigilance, while common in human-machine interaction paradigms, are not inherent to AI-assisted driving (also referred to as "Level 2", "semi-autonomous", and "partially automated" driving). One implication of this is that it may be possible to design AI-assisted vehicles that rely on humans for supervision in a way that will not necessarily lead to over-trust and significant vigilance decrement. We propose a measure of "functional vigilance" that conceptualizes vigilance when drivers are allowed to self-regulate by choosing when and where to leverage the capabilities of automation and when to perform the driving task manually. The central observations in the dataset is that drivers use Autopilot for 34.8% of their driven miles, and yet appear to maintain a relatively high degree of functional vigilance. These observations are based on annotation of 18,928 disengagements of Autopilot that quantify the ability of drivers to respond to challenging driving situations during AI-assisted driving. We discuss limitations and implications of this work including that these findings (1) cannot be directly used to infer safety as a much larger dataset would be required for crash-based statistical analysis of risk, (2) may not be generalizable to a population of drivers nor Autopilot versions outside our dataset, (3) do not include challenging scenarios that did not lead to Autopilot disengagement, (4) are based on human-annotation of critical signals, and (5) do not imply that driver attention management systems are not potentially highly beneficial additions to the functional vigilance framework for the purpose of encouraging the driver to remain appropriately attentive to the road. The authors are highly cognizant that there are significant nuances in the design, analysis, and interpretation of this work. It is our hope that it will encourage serious discussion and further investigation of how seemingly subtle features of AI-assisted system design and implementation may influence the extent to which humans are able to sustain appropriate collaborative engagement with such technology.**



(a) Classical vigilance framework.

(b) Functional vigilance framework.

Fig. 1: Two approaches to measuring vigilance of human-in-the-loop automation systems. The critical distinction is that in the classical approach, the operator does not have the option to choose when and where to engage in the monitoring task and instead must always be in the supervisory role. On the other hand, in the functional approach, the operator can choose to switch between task performance and machine monitoring at any time.

* Corresponding author: fridman@mit.edu

I. Introduction

The twenty first century may very well be defined by global proliferation of artificial intelligence and robotics systems that directly interact with and are responsible for the well-being of human beings. Successful utilization of AI and robotics may venture out of well-lit, well-controlled factory floors and the digital constraints of cyberspace, and enter at a mass-scale into our homes (i.e., personal robotics), our skies (i.e., delivery drones), and our transportation network (i.e., autonomous vehicles). What is not well understood about this new world is how ready the underlying algorithms are to perceive, reason about, and act in the world they enter. No more is this true than with autonomous vehicles, where each artificial intelligence system is directly responsible for the comfort and safety of vehicle occupants and surrounding road users. The question, therefore, is how can that integration of AI into society save lives, not take them, while enhancing the overall driving experience. Central to this question is the understanding of human behavior in the context of AI-assisted driving, measured by the ability of the driver to maintain vigilance, to gain understanding of system limitations, and to balance trust and risk in a world full of uncertainty where a split-second gap in judgment may be catastrophic.

The focus of this work is Tesla Autopilot, a Level 2 [1] driving system capable of automated longitudinal and lateral control under supervision of the human driver. Taxonomization of automation capabilities have received significant efforts in recent years, resulting in a number of terms used for systems with capabilities similar to Autopilot. In the popular press and in robotics, machine learning, and AI communities the most prevalent term is "semi-autonomous" systems [2]. In automotive standard bodies and human factors communities, the most prevalent term is "partially automated" [3, 4]. The Society of Automotive Engineers (SAE) refers to Level 2 systems as "partial driving automation" [1]. Some prominent activists, members of industry, and vehicle assessment agencies (e.g., Euro NCAP) argue that the use of words such as "autonomous" and "automated" as part of the terminology (or the term "Autopilot" itself) is irresponsible as it leads to consumer misunderstanding of the limitations of the technology [5, 6]. Research on a range of actual and invented feature names, including Autopilot, has shown a connection between the phrase "auto" and expectations of capabilities exceeding that of the actual technologies [7]. In order to avoid misinterpretation of system capabilities under investigation in this work, we avoid the words "autonomous" and "automated" and instead use the term "AI-assisted" driving that clarifies that Autopilot is assistive not fully autonomous technology, and thus requires human supervision at all times.

Due to its scale of deployment and individual utilization, Autopilot serves as perhaps the currently best available opportunity to study and understand human interaction with AI-assisted vehicles "in the wild." Tesla Autopilot is a system that uses a mixture of vision, radar, ultrasonic, and GPS sensors as input to control both longitudinal and lateral movement.

To date, Tesla vehicles have driven over 1 billion miles under the control of Autopilot [8] since its activation on Tesla vehicles in October 2015. And yet, even with so many miles traveled, we know very little about when, where, and how the human supervisors elect to utilize the automation capabilities of the system. A few anecdotal accounts published on social media, often for the purpose of entertainment, and small scale studies [9, 10] have begun to detail perceptions of and actual system use and behavior. However, such small-scale experimental studies and limited observational efforts may not provide generalizable insights into human behavior in the naturalistic driving context.

We aim to move away from anecdotal reports and toward more objective, representative analysis of real-word use. While the ultimate test of an autonomous or AI-assisted driving system will be a statistical consideration of safety over tens or hundreds of billions of miles traveled [11], naturalistic driving research can now begin investigating and identify both promising and concerning trends in drivers' behavioral patterns in the context of Autopilot.

To this end, we propose a measure of driver "functional vigilance" that incorporates both the ability of the driver to detect critical signals and the implicit ability of the driver to self-regulate when and where to switch from the role of manually performing the task to the role of supervising the automation as it performs the task. See Fig. 1 for a diagram contrasting classical and functional vigilance frameworks, and §III-A for a detailed discussion of the concept. In the context of vehicle automation that evaluates the driver's ability to respond to challenging driving situations, we focus on "tricky situations", a term that refers to scenarios that, if not attended to, may lead to property damage, injury, or death. We use this measure to evaluate 8,682 Autopilot disengagements in response to tricky situations, and show that drivers in our dataset remain vigilant (as defined below) when they elect to use Autopilot. They choose to do so for 34.8% of total miles driven. This choice of when and where to use Autopilot is the additional element that highlights the difference between the classical study of vigilance and a "functional" study of vigilance. Functional vigilance incorporates both the ability of the driver to detect critical signals and the implicit ability of the driver to be strategic about when and where to use Autopilot.

This work takes an objective, data-driven approach toward evaluating functional vigilance in real-world AI-assisted driving by analyzing naturalistic driving data in Tesla vehicles that were instrumented for data collection as part of the MIT Autonomous Vehicle Technology Study (MIT-AVT) [12].

The two main results of this work are that (1) drivers elect to use Autopilot for a significant percent of their driven miles and (2) drivers do not appear to over-trust the system to a degree that results in significant functional vigilance degradation in their supervisory role of system operation. In short, in our data, drivers use Autopilot frequently and remain functionally vigilant in their use. The quantitative summary of the latter is presented in Table I.

These results are surprising as they do not align with the prediction of prior literature on human monitoring of automation, which predicts significant degradation of vigilance as covered in §II. In §III-B, we describe when and how Autopilot is used by drivers in our study. In §III-A, we define the measure of functional vigilance and report this measure for 15,874 human-initiated and 128 machine-initiated disengagements of Autopilot. Finally, in §V, we present limitations of the work, and provide a perspective that the imperfection of the system is counter-intuitively central to maintaining functional vigilance and limiting over-trust of the system.

Finally, given the potentially impactful nature of the findings, it is important to emphasize, as stated previously, that these findings (1) cannot be directly used to infer safety as a much larger dataset would be required for crash-based statistical analysis of risk, (2) may not be generalizable to a population of drivers nor Autopilot versions outside our dataset, (3) do not include challenging scenarios that did not lead to Autopilot disengagement, (4) are based on human-annotation of critical signals, and (5) do not imply that driver attention management systems are not potentially highly beneficial additions to the functional vigilance framework for the purpose of encouraging the driver to remain appropriately attentive to the road.

II. Related Work

Human vigilance in the context of monitoring automation has been studied for over 70 years [13]. The central measure of interest is the "vigilance decrement" which is the decrease in a human being's ability to remain vigilant for critical signals over time, as indicated by a decline in the rate of the correct detection of signals [14]. Taken together, this body of work forms a definitive set of findings which identify the conditions under which human over-trust automation systems that are highly reliable, lose vigilance during supervision of these systems, and fail to detect safety-critical events at rates higher than when performing the task themselves. These findings form a foundation from which to build our understanding of vigilance in the context of AI-assisted driving. In this section, we review select key findings from three areas of research: (1) fundamental vigilance concepts and studies, (2) vigilance in aviation, and (3) vigilance studies in driving simulators. This review of the literature provides a foundation upon which our framework of functional vigilance as described in §III-A is developed.

A. Fundamental Vigilance Concepts and Studies

The bulk of the work on vigilance decrement over the past 70 years has been conducted under controlled conditions such that individual variables affecting vigilance could be rigorously studied. These experiments were conducted both in the laboratory and in the field. Applications span a wide range of domains from general visual search tasks [13] to medical diagnosis tasks [15] to agriculture [16]. The majority of results across studies observe that humans make errors on tasks that require prolonged attention. These observations carry over to the task of monitoring and supervising automation, finding that increases in level of automation correlate with an increase in vigilance decrement [17, 18].

Typically, within automation, the human supervisor is expected to monitor the automation for failures in its performance that would require the human operator to intervene or take action of some type. Within the driving context, however, the situation is more complicated. For AI-assisted driving at SAE Level 1 (L1) or Level 2 (L2), the human supervisor needs to monitor automation status and performance as shown in Fig. 1. In addition, they need to monitor the driving scene because object and event detection remains the responsibility of the human at these levels of automation. In real-world driving, the process of "monitoring" is not as simple as looking at a sequence of novel images and attempting to detect an anomaly. Monitoring involves integrating spatial and temporal context from all the senses and across time from seconds to years, including prior interaction with the current driving scene and driving scenes similar to it in some semantic or actionable way.

Research in the scientific literature has shown that highly reliable automation systems can lead to a state of "automation complacency" in which the human operator becomes satisfied that the automation is competent and is controlling the vehicle satisfactorily. And under such a circumstance, the human operator's belief about system competence may lead them to become complacent about their own supervisory responsibilities and may, in fact, lead them to believe that their supervision of the system or environment is not necessary. This can, in turn, mean that the quality of human monitoring may decline below some standard level of performance and may lead to missed system failures or delayed responses to a system failure [19]. In the context of driving, this can additionally include missed or delayed responses to objects and events in the driving environment (e.g., missed detections of unexpected pedestrians or bicyclists intruding into the vehicle's path of travel; missed detections of unexpected braking of a lead vehicle, missed detections of unexpected behavior of vehicles in adjacent lanes, etc.)

The extent of degradation in critical event detection that occurs under complacency is affected by several factors including the reliability of system automation and the total workload that is being carried by the human operator at the time. Parasuraman et al. [20] found that the mean detection rate of automation failures was markedly higher under conditions where the reliability of the system varied (82%) vs. under conditions where the system's reliability was constant (33%). Additionally, complacency and vigilance decrements are more often found under conditions where task load is high. This is particularly true where multitasking is present, and thus there are competing tasks to which the human operator's attention may be allocated when they become complacent about the automation.

The corollary to increased complacency with highly reliable automation systems is that decreases in automation reliability should reduce automation complacency, that is, increase the detection rate of automation failures. Bagheri and Jamieson [21] did find that participants detected significantly more automation failures at low rather than at high automation reliability. We explore this concept as a possible explanation for the observed results in §V-A2.

In the foundational studies on automation complacency done by Parasuraman and colleagues, the rate of automation failures which humans were trying to detect was relatively high – for example, 12% in the "high" reliability condition [20] which corresponds to a rate of about 1 every 3.5 minutes. However, Parasuraman and Manzey [19] described this as a drawback of studies published in the literature, because they felt that such high failure rates were unrepresentative of any real automated system or at least unrepresentative of any system that human operators would use.

Wickens & Dixon [22] hypothesized that when the reliability level of an automated system falls below some limit (which the suggested lies at approximately 70% with a standard error of 14%) most human operators would no longer be inclined to rely on it. However, they reported that some humans do continue to rely on such automated systems. Further, May [23] also found that participants continued to show complacency effects even at low automation reliability. This type of research has led to the recognition that additional factors like first failure, the temporal sequence of failures, and the time between failures may all be important in addition to the basic rate of failure.

B. *Vigilance in Real-World Aviation*

The focus of our work is AI-assisted driving in the real world. This problem has not been extensively studied to date, except in the simulator (see §II-C). The closest domain of human-machine interaction that has been studied in the real-world is automation in aviation. Much like with driving, the bulk of the work in aviation is in the simulator [24]. However, observational reports on real-world general aviation have been published [25]. The findings generally suggest that the vigilance decrement and human tendency to over-trust is ubiquitous. However, such decrement can be alleviated in part through a number of countermeasures including training, regular briefings, effective communication with the crew, review and modification of plans [26].

C. *Vigilance in Driving Simulators*

A large number of studies of AI-assisted and fully autonomous driving have been conducted in driving simulators [27, 28]. Many have observed a vigilance decrement as measured by the driver's ability to respond to challenging situations. Greenlee et al. [29] showed that in a 40 minute automated drive in a simulator, hazard detection rate declined precipitously, and reaction times slowed as the drive progressed. Carsten et al. [30] showed that in a 45 minute automated drive in a simulator, drivers shifted attention away from the driving task and tended to use the automation support to enable engagement in nondriving tasks. Automated driving in these studies refers to longitudinal and lateral control akin to the capabilities of Autopilot.

Several dozen papers of such experiments in driving simulators have been published over the past 10 years [31]. Several of these studies explore what factors contribute to inattention and vigilance decrement in the driving simulator, and many find that there is in fact a significant vigilance decrement. The degree to which these studies of automation generalize to the real world is unknown [32].

To move beyond this limitation, we examine the functional vigilance of drivers operating their Tesla vehicles in the wild, as they naturally drive. Data acquired from their driving allows us to examine tricky situations and critical events that arise, identify those leading to transfers of control, identify the rate of their occurrence, and the way that they are handled in order to understand whether or not there are signs of automation complacency or degradations of functional vigilance in the behavior of drivers.

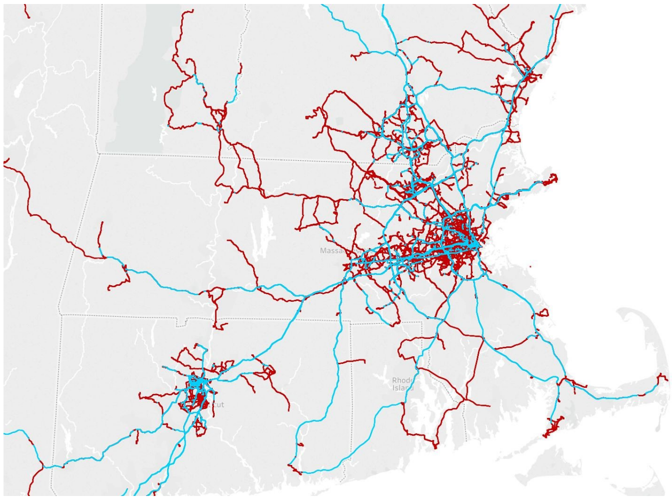


Fig. 2: Visualization of the Tesla vehicle data collected in MIT-AVT study. Red lines designate manual control of the vehicle. Cyan lines designate Autopilot control of the vehicles.

III. Methods

A. *Functional Vigilance*

As discussed in §II-A, vigilance is classically defined as the ability of a human being to maintain concentrated attention on a task that requires the detection of a critical signal. In the automation context, the vigilance decrement measures the decrease in a human ability to detect critical events when tasked with supervising the automation. This classical vigilance framework is shown in Fig. 1a.

Building upon the basic definition of vigilance (allocating attention to detection of critical events over a period of time), we wish to introduce a variation on it: the concept of "functional vigilance."

We do not deviate from this definition of vigilance in measuring the fundamental ability of a driver to detect critical events. However, we introduce a new term of "functional vigilance" to emphasize the methodology and framework within which we measure driver vigilance. The central characteristic underlying the concept of functional vigilance is the ability of the driver to choose when to serve as the operator of the vehicle and when to serve as the supervisor of the automation (in this case, Autopilot). Majority of the work on vigilance decrement over the past several decades (see §II) does not allow for this choice. In driving simulator studies, the usual experiment has the human supervise the machine as it operates for a specific period of time under a constant level of automation. The free choice of when, where, and how to serve as the supervisor is not given to the driver in most cases. The ability to make this choice, however, may be the critical pre-requisite of successful self-regulation of vigilance.

There is an important and illuminating distinction between functional vigilance and "driver focus." The latter is a general measure of the degree that the driver is paying attention to the driving scene [33]. Functional vigilance measures the ability of the driver to detect and respond to critical events when they arise. In AI-assisted driving, the two measures may be highly correlated or they may not be. This is an important open question that is not addressed in this work.

B. AI-assisted Driving Dataset

The dataset used in this work, which we refer to hereinafter as the "Autopilot dataset", is all the data from Autopilot-enabled Tesla vehicles that are part of the MIT Autonomous Vehicle Technology (MIT-AVT) naturalistic driving study [12]. The purpose of the MIT-AVT study is to collect and analyze large-scale naturalistic data of AI-assisted driving in order to understand and characterize real-world interaction between human drivers and autonomous driving technology.

The Autopilot dataset includes 21 Autopilot-capable Tesla vehicles and 323,384 total miles. The Tesla vehicles are all driver-owned. No restrictions, suggestions or other guidance is placed on where, when and how the vehicles are driven. As shown in Fig. 2, the bulk of the driving in the dataset is located in the Greater Boston and New England region, although extended drives (e.g., from Massachusetts to Florida and to California) are present in the dataset.

To observe, annotate, and automatically sense the driver and the driving scene, the vehicles are instrumented with three cameras: (1) on the driver's face, (2) on the in-cab region, and (3) directed out at the forward roadway [12]. All three camera streams are watched jointly in synchrony during the manual annotation process as described in §IV-B. Autopilot-related system state, vehicle kinematics, and other pertinent signals are derived from messages on one of the vehicle CAN buses.

The Autopilot dataset contains a total of 26,638 epochs of Autopilot utilization. An Autopilot epoch is defined as a period of time between the driver electing to engage Autopilot and either the driver or the system itself disengaging it. The focus of the vigilance analysis in this work is on an epoch of time before and after Autopilot disengagement (see §IV-B).

C. Autopilot Dataset System Specification

Tesla vehicles include several advanced safety and driver assistance features. Under consideration in this work are the modules of Traffic Aware Cruise Control (TACC) and Autosteer, providing adaptive cruise control and lane-centering capabilities, respectively. From this perspective, Tesla can operate in 3 distinct modes: (1) manual control, (2) only TACC, and (3) Autopilot (both TACC and Autosteer).

There are two hardware versions of Autopilot [8] in the dataset, termed hardware version 1 (HW1) and hardware version 2 (HW2). HW1 includes sensors fusion of radar, ultrasonic sensors, and a monocular camera system developed by Mobileye. HW2 includes eight surround cameras that provide 360 degrees of visibility around the vehicle at up to 250 meters of range, twelve updated ultrasonic sensors, and a forward-facing radar. Of the 21 vehicles in the dataset, 16 are HW1 vehicles and 5 are HW2 vehicles.

D. Transfer of Control

The transition between the three aforementioned modes is well-defined and requires explicit action by the driver, except for machine-initiated disengagements of Autopilot that are accompanied by a loud audible warning and the visual displayed symbol of the colloquially named "red hands." In this work, we do not consider special cases of automatic emergency braking (AEB), automatic lane change, and "navigate on Autopilot" capabilities.

In Tesla Model S and Model X vehicles, the Autopilot stalk is located to the left of and slightly behind the steering wheel. Autopilot is engaged by pulling this stalk twice when in manual state and once when in TACC state. An icon of a steering wheel on the right side on the instrument cluster is the main indicator of Autopilot state. The icon is (a) blue when Autopilot is engaged, (b) gray when it is available but not engaged, and (c) not visible when Autopilot is not available.

Autopilot can be disengaged in 4 ways: one system-initiated and three driver-initiated. In a system-initiated disengagement, Autopilot provides a visual "red hands" cue and an auditory cue that indicates to the driver that they must immediately take control of the vehicle. The three driver-initiated disengagement options are via braking, steering, or pushing the stalk.

E. Critical Events during Autopilot Driving

Measuring functional vigilance requires enumerating categories of critical events (CE) in Autopilot driving. Four categories were used in this study: CE1, CE2, CE3, and CE4. These are defined below. The term "tricky situations" is used in the definitions and throughout this work to describe challenging driving scenarios that require a response or anticipatory action by the driver in order to maintain safe operation of the vehicle. The four categories are defined as follows:

1) **CE1:** Human-initiated disengagements of Autopilot in anticipation of or in response to tricky situations.
2) **CE2:** System-initiated disengagement of Autopilot associated with tricky situations.

3) **CE3:** Sudden deceleration events (i.e., hard braking) during Autopilot control.
4) **CE4:** Tricky situations during Autopilot control that do not result in disengagement or crash.

CE1 and CE2 events are a subset of all Autopilot disengagement events annotated as being associated with a tricky situation. We describe this annotation process in §III-F. However, a prerequisite step to this annotation is the automated discovery of Autopilot disengagement events. These disengagements were extracted from the dataset by monitoring Autopilot state CAN bus messages and detecting the moments when the state changes from Autopilot enabled to any other state. In addition, the same state variable provided error values associated with system-initiated disengagements, allowing us to automatically label disengagement as human-initiated or machine-initiated. There are a total of 26,638 disengagement epochs in the dataset considered here. We filtered out a set of epochs that were difficult to annotate accurately. This set consisted of disengagements (1) when Autopilot was used for less than 5 seconds and (2) the sun was below the horizon computed based on the location of the vehicles and the current date. Clear visibility of both the driving scene and the driver are paramount for the annotation process that categorizes the reasons for the disengagements. 18,928 disengagement epochs resulted from this filtering process and went to the annotation process as described in §III-F. Disengagement epochs resulting in a crash (defined as striking another solid object) fall into CE1 and CE2. No such crashes were detected in our dataset.

CE3 epochs were detected based on a 0.6g deceleration trigger. This trigger type and threshold value was found to be the most effective criteria for detecting crash-relevant events in the SHRP2 dataset [34]. While 278 sudden deceleration events were detected during manual control of the vehicle, zero such events were detected during Autopilot control. It is not precisely known what the maximum braking deceleration threshold for the vehicles in the study is before the automatic emergency braking (AEB) system is triggered. If the threshold is below 0.6g, then the absence of CE3 epochs is part of the system design specification, and harder braking events would trigger AEB events and system disengagement. Braking events that took the vehicle out of Autopilot control fall into CE1 events, and are analyzed as part of the disengagement annotation process in §III-F. Therefore, any hard braking events in the 10 second window following the disengagements are annotated for their timeliness of driver response (see §III-F).

CE4 events are critical events that happen during Autopilot but do not result in a disengagement or crash. Examples may include running a red light while on Autopilot or drifting across multiple lanes without the driver having initiated an automatic lane change. It is hypothesized that events of this type are most often captured in the CE1 and CE2 categories, because they would most likely result in a disengagement or crash. However, it is possible to imagine cases where such critical events do not result in a disengagement, but instead result in Autopilot regaining a safe and proper trajectory on the road. Such events are difficult to discover in the data – and it is not clear that they can be automatically discovered using computer vision, particularly without extensive supervised machine learning efforts targeted at discovering each specific rare edge case based on prior knowledge of each case's visual and semantic characteristics. We discuss possible computer vision methodologies for discovering such events in §VI and the limitations associated with excluding these events in §V-B.

However, it is very important to understand that the CE4 events may be very complex in nature – and while every effort was made to find CE4 events in this research, it is still possible that a few of these very rare events went undetected in the dataset. Furthermore, it is quite possible that CE4 events are categorically different from CE1-3 – and require not only that the driver be attentive to how well the AI-assisted system is performing, but also require the driver to be attentive to the driving situation and the degree to which the system's performance is appropriate to that situation. This may requires in-depth knowledge of the system and its limits. It may require more than simple "detection of visual stimuli". In other words, it may require more than vigilance. It may require in-depth understanding of the system [35, 36] and the ability to diagnose the situation. In fact, it can be argued that identifying events in CE4 (at least some of them) cannot be treated within the framework of vigilance, because they may involve much more complex issues of knowledge-driven diagnosis, decision, and response. For example, when an Autopilot-enabled Tesla has been following a lead-vehicle (LV) and the LV leaves the lane – revealing a stopped object (e.g., a fire engine) – wherein the driver would need to understand that the Autopilot system may not "see" stopped objects in this circumstance, nor brake to them. Another example is when Autopilot treats a "gore" (road split or exit ramp) in the highway as a set of lane lines, the driver would need to understand and recognize that this has happened and intervene before the gore ends in a barrier.

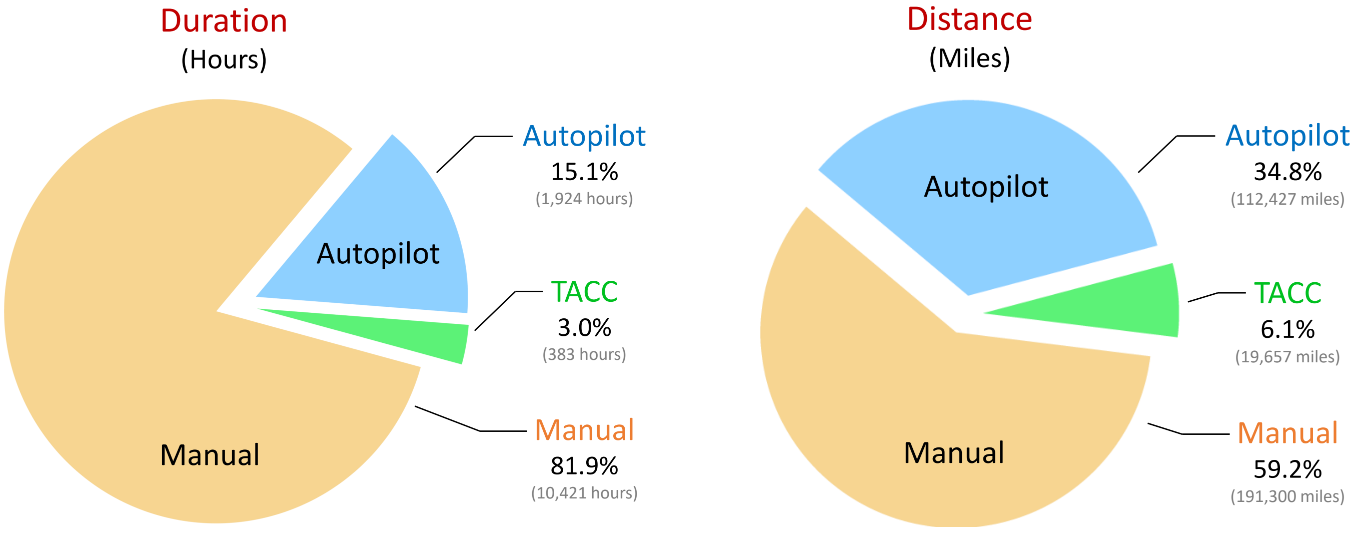
F. Annotation of Autopilot Disengagement Epochs

Each epoch of the 18,928 Autopilot disengagements epochs were manually annotated by 3 to 13 people, depending on the level of a disagreement between the annotators. The clips show 5 seconds before and 10 seconds after the Autopilot disengagement event.

The question asked of the annotators was: *"Why was Autopilot disengaged in this situation?"* Supplementary information provided along with the question was as follows:

- A tricky situation is a challenging driving scenario that requires anticipatory or responsive action by the driver in order to avoid potential property damage or a crash.
- If a tricky situation is present, use one of the three responses associated with tricky situations.
- Examples of tricky situations include approaching a sharp curve, lane merging, drifting out of lane, moving too close to road dividers or other cars, vehicle or pedestrian blocking road, etc.

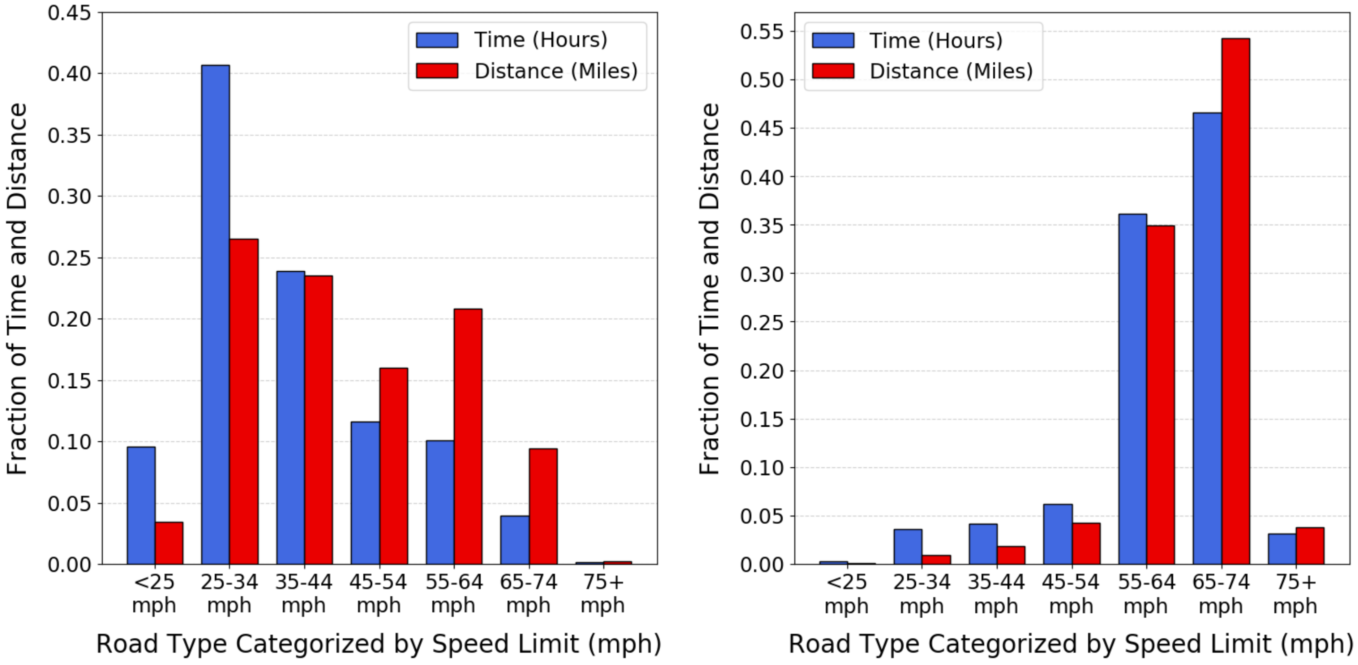
For human-initiated disengagements, the answer options were those listed in Table I excluding "hands off wheel." For



(a) Fraction of time traveled by automation mode.

(b) Fraction of distance traveled by automation mode.

Fig. 3: Fraction of time (hours) and distance (miles) traveled under manual, TACC, and Autopilot control.



(a) Vehicle under manual control.

(b) Vehicle under Autopilot control.

Fig. 4: Time spent and distance traveled on different road types categorized by speed limit.

machine-initiated disengagements, the answer options were more extensive but mapped directly into the list provided in Table I.

We used a majority vote criteria for the annotation mediation process. The number of annotations performed on each epoch started at 3 and increased in increments of 2 until more than 50% of the annotators selected the same label for a clip. 13 annotators was the maximum number reached in order to achieve a majority.

G. *Subjective Nature of Annotating Response Timeliness*

Our formulation and terminology of "tricky situations" emphasizes the subjective nature of characterizing the critical signals based on which functional vigilance is evaluated. There are three subjective elements to these situations:

1) **Planned vs Unplanned:** The degree to which a short-term navigation decision is planned (and thus not tricky) or unplanned (and thus tricky).
2) **Timing of Critical Signal:** When does the tricky situa-

tion arise based on which a response of the driver may be warranted.

3) **Timing of Driver Response:** What is the gap in time between when the critical signal is reasonably expected to be detected and when the driver takes observable action to successfully respond to the signal. The first choice is whether the action is anticipatory or responsive (before and after critical signal, respectively). The second choice is whether a responsive action is "immediate" or "too late".

The nature of "too late" in evaluating the timeliness of a response is a difficult one to characterize precisely. At one extreme of defining "too late" are avoidable scenarios that lead to a crash. At the other extreme is any scenario where a tricky situation could have been addressed earlier. We provide a definition of "too late" in order to aid the subjective annotation. The definition is as follows: *Situations where the driver both should have and could have responded in a more timely manner in order to avoid safety-critical consequences of a tricky situation.*

Additionally, we added a guiding note: *Any response delayed by more than 1 second after the tricky situation should likely be annotated as "too late."*

IV. Results

A. Patterns of Use

As shown in Fig. 3, the Autopilot dataset includes 323,384 total miles and 112,427 miles under Autopilot control. Autopilot is used to drive 34.8% of miles and 15.1% of hours as shown in Fig. 3. This represents a significant use of Autopilot and reflects that the drivers are deriving value from the use of the system. In contrast, TACC alone is only used 3% of the time. Consequently, for the majority of our analysis, we do not consider TACC, and only focus on the comparison and transition between manual and Autopilot, which together comprise 97% of driving time.

Fig. 4 shows the time spent and distance traveled under manual and autopilot control on different road types. Speed limit, derived from the fusion of GPS and vision sensors, is used for categorizing road type following the taxonomy defined by the U.S. Department of Transportation Federal Highway Administration [37]. Interstate, freeway, multilane highway, and other arterial roads are generally associated with speed limits of 55 mph and above. Connector roads and local roads are associated with speed limits below 55 mph. Drivers elect to use Autopilot primarily on roads with a speed limit of 55 mph and above. In contrast, the vehicle is controlled manually on roads with speed limits below 55 mph the majority of the time. This is true both when measured in time and distance spent on these road types.

To further contextualize patterns of system use, we report on the fraction of time spent in speed-restricted traffic, defined as travel speeds 10 mph below the speed limit or slower. Measured in time, 45.5% of manual control, 15.87% of TACC control, and 19.3% of Autopilot control is spent in these kinds of traffic conditions. In other words, in our dataset, Autopilot is primarily used in fast, free flowing traffic as measured by both fraction of time and distance.

Normalizing to the number of Autopilot miles driven during the day in our dataset, it is possible to determine the rate of tricky disengagements. This rate is, on average, one tricky disengagement every 9.2 miles of Autopilot driving. Recall that, in the research literature (see §II-A), rates of automation anomalies that are studied in the lab or simulator are often artificially increased in order to obtain more data faster [19] such as "1 anomaly every 3.5 minutes" or "1 anomaly every 30 minutes." This contrasts with rates of "real systems in the world" where anomalies and failures can occur at much lower rates (once every 2 weeks, or even much more rare than that). The rate of disengagement observed thus far in our study suggests that the current Autopilot system is still in an early state, where it still has imperfections and this level of reliability plays a role in determining trust and human operator levels of functional vigilance. We discuss this concept as a possible explanation for the observed functional vigilance results in §V-A2.

B. Functional Vigilance during Autopilot Driving

The measure of functional vigilance we use in our analysis is not merely whether the drivers detect the critical events CE1 and CE2 defined in §III-E but if they do so in a timely fashion. The three temporally distinct categories associated with tricky situations during Autopilot disengagement are shown in Table I under the category of "Tricky Situation Present." The first two subcategories when the driver performs anticipatory action or responds immediately to a tricky situation are indicative of a high level of functional vigilance. The third subcategory when the driver responds too late to a tricky situation would be the category which would include instances arising from a low level of functional vigilance.

Of the 18,928 annotated disengagement epochs, 8,729 epochs were labeled as associated with tricky situations. Their description and distribution is listed under "Tricky Situation Present" category in Table I. The target measure for this analysis is the number of epochs annotated as "act too late after tricky situation." These epochs are those that would be considered missed or delayed detections of critical events and thereby would represent a significant functional vigilance decrement. As Table I shows, no such epochs were discovered in our dataset. The high-level functional vigilance breakdown of Autopilot disengagement epochs is as follows:

- No tricky situation: **10,118**
- Tricky situations that were anticipated ahead of time or responded to immediately: **8,682**
- Tricky situations that were responded to after a significantly delay or not at all (see note in Table I): **0**

Table II shows the categories of tricky situations and their frequencies for disengagements that preceded and followed a tricky situation. Presence of a curve is the most common reason for anticipatory disengagement of autopilot. The vehicle getting too close to lane, wall, or another car, is the most

Critical Event Category	Disengagement Reason	Description	Human Initiated	Machine Initiated
Tricky Situation Present	Act too late after tricky situation	Delayed response to tricky situation (see details in §III-G).	0	0
	Act right after tricky situation	Rapid timely response after a tricky situation arises.	813	47
	Act before tricky situation	Anticipatory action before a tricky situation.	7,869	0
No Tricky Situation Present	Planned Turning or Speed Change	Taking control to make a planned navigation decision.	8,608	68
	Planned Stopping	Stopping for stop sign, yellow/red traffic light.	601	0
	Accidental	Accidentally bumping the wheel or the Autopilot stalk.	38	0
	Annotation Difficult	Image is too bright/dark for accurate annotation.	94	0
	No clear reason	No clearly identifiable reason	777	0
	Hands off wheel	Warning ignored while remaining attentive to the road.	0	13
		Total Annotated Disengagement Epochs:	**18,800**	**128**

Summary of Results:

Type of Driver Response	Percentage of Disengagements
Delayed (Slow responses or missed detections)	0.0%*
Responsive (Rapid timely responses)	4.5%
Anticipatory (Action before T.S. or planned decision)	90.6%
Other (Accidental, not annotatable, etc.)	4.9%

*This value is entered as "0.0%" to reflect the fact that no such events were found in our dataset using the methods described. However, it is possible that some events of this type exist in the dataset but went undiscovered. Future work may lead to new methods that will help identify these, if any exist.

TABLE I: Annotated reasons for disengagement of Autopilot. The annotation process and question details are described in §III-F. Reasons are divided into two categories: those associated with tricky situations and those that are not. The label of "act too late after tricky situation" was designed to locate disengagement epochs associated with high functional vigilance decrement. Of the 18,928 total annotated disengagements, no epochs were labeled in this way by the annotators. The results are summarized in the table on the right with respect to functional vigilance and anticipatory characteristics of the quantitative results.

common reason for reactive disengagement of Autopilot. This statistics may be instructive for designing AI-assisted systems that aim to effectively deal with scenarios that drivers may consider "tricky."

Fig. 5 shows the difference in disengagement velocity and epoch duration between Autopilot epochs that do not end in a tricky situation and those that do, the latter receiving either an anticipatory or a reactive action from the driver. A key statistic is that the median duration for an Autopilot epoch that ends in a reactive response to a tricky situation is 2.0 minutes and the mean is 5.2 minutes. The distribution of epoch duration has a very long tail including many epochs that are over 1 hour. In general, all metrics considered revealed no significant differences between the three categories of Autopilot disengagements shown in Fig. 5.

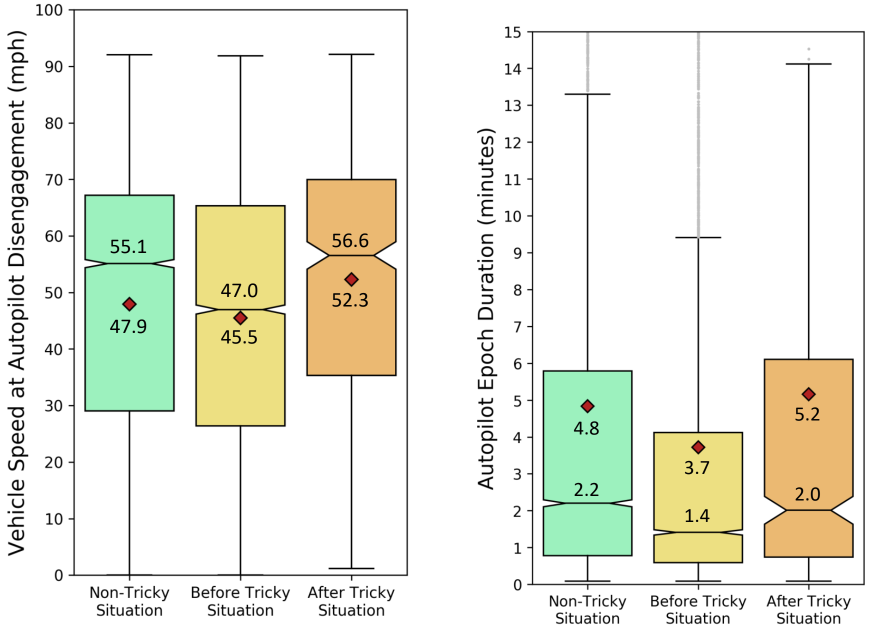
V. Discussion

A. *Proposed Explanation for Observed Behavior*

The patterns of Autopilot use and the functional vigilance measures reported in this work indicate that drivers in this study were using Autopilot extensively and yet did not appear to over-trust the system to a degree that compromised functional vigilance. We hypothesize two explanations for the results as detailed below: (1) exploration and (2) imperfection. The latter may very well be the critical contributor to the observed behavior. Drivers in our dataset were addressing tricky situations at the rate of 1 every 9.2 miles. This rate led to a level of functional vigilance in which drivers were anticipating when and where a tricky situation would arise or a disengagement was necessary 90.6% of the time. In another 4.5% of cases, drivers were immediately responsive to an Autopilot disengagement or tricky situation, suggesting that they were attentive and functionally vigilant.

It is important to emphasize, as stated previously and detailed further in §V-B, that these findings (1) cannot be directly used to infer safety as a much larger dataset would be required for crash-based statistical analysis of risk, (2) may not be generalizable to a population of drivers nor Autopilot versions outside our dataset, (3) do not include challenging scenarios that did not lead to Autopilot disengagement, (4) are based on human-annotation of critical signals, and (5) do not imply that driver attention management systems are not potentially highly beneficial additions to the functional vigilance framework for the purpose of encouraging the driver to remain appropriately attentive to the road.

1) Exploration: Most of the time and distance traveled under Autopilot control in our dataset is in free-flowing traffic on highways. However, a significant fraction of Autopilot epochs and thus disengagements are on local roads (i.e., roads characterized by non-highway speed limits). This may indicate that drivers regularly explore the limits of the system in a way that ventures outside the traditionally defined operational



Max	92.1	91.9	92.2
Q3	67.2	65.3	70.0
Median	55.1	47.0	56.6
Mean	47.9	45.5	52.3
Q1	29.1	26.4	35.3
Min	0.0	0.0	1.2

(a) Vehicle speed at the moment of Autopilot disengagement.

Max	126.6	68.0	73.0
Q3	5.8	4.1	6.1
Mean	4.8	3.7	5.2
Median	2.2	1.4	2.0
Q1	0.8	0.6	0.7
Min	0.1	0.1	0.1

(b) Autopilot epoch duration.

Autopilot Epoch Duration	Number of Autopilot Epochs
>0 mins	18,928
>5 mins	4,914
>10 mins	2,332
>15 mins	1,270
>20 mins	676
>25 mins	400
>30 mins	224
>35 mins	126
>40 mins	86
>45 mins	56
>50 mins	37
>55 mins	30
>60 mins	16

(c) Number of Autopilot epochs by duration lowerbound.

Fig. 5: Autopilot epoch statistics categorized by presence and timing of tricky situation. The boxplots and the accompanying tables in (a) and (b) show the summary statistics of the underlying epochs. The table in (c) shows the number of Autopilot epochs longer than a specific duration.

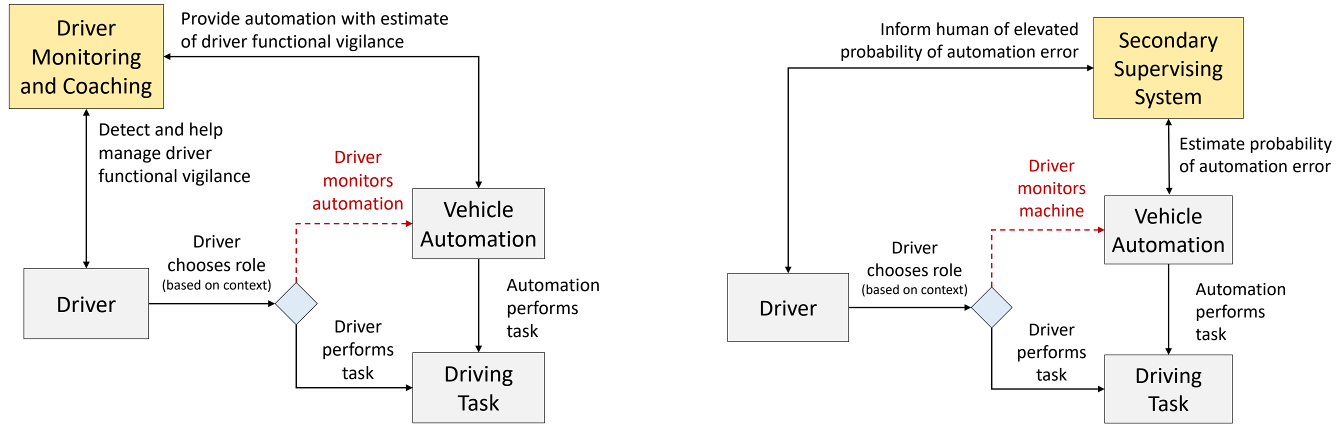
design domains (ODD) for similar vision-based lane-centering systems. This type of experiential learning may allow them to acquire an understanding of system performance limits – knowledge which is then used in dealing with events like tricky situations described in this research. At this time, this hypothesis is based on discussion with Tesla owners and preliminary results of survey responses from Tesla owners. In future work, we will seek to support or disprove this hypothesis through methods which may include extensive self-report data collection and further data analysis.

2) Imperfection: In common with most, if not all, current AI-assisted vehicle systems, Autopilot is not perfect in its ability to safely navigate through any possible edge case scenario in driving. In our data, as described in §IV-B, 46.2% of Autopilot disengagements are where the human driver is anticipating or responding to a tricky situation. Normalizing this number of Autopilot miles driven during the day in our dataset, we determine that such tricky disengagement occur on average every 9.2 miles of Autopilot driving. Under these conditions, the system limits reveal themselves regularly and the human driver “catches” the system and takes over. The natural engineering response to such data may be to strive to lower the rate of such “failures.” And yet, these imperfections are likely a significant contributing factor to why the drivers are maintaining functional vigilance. In other words, *perfect may be the enemy of good* when the human factor is considered. A successful AI-assisted system may not be one that is 99.99...% perfect but one that is far from perfect and effectively communicates its imperfections.

B. Limitations

The central finding of this analysis of our large-scale naturalistic AI-assisted driving study is that drivers in this sample operating Tesla vehicles under Autopilot control appear to remain functionally vigilant to a degree that stands in contrast to what might be predicted by prior literature that spans fields from robotics [39] to human factors [17]. While, to the best of our knowledge, this study is the largest published work of its kind, and its findings are grounded in real-world data, we acknowledge possible limitations and anticipate differing interpretations of the data. In this section, we present several such limitations.

1) Subject Sample Characteristics and Demographic Generalizability: The most common concern raised in our discus-


Driver Monitoring and Coaching
Provide automation with estimate of driver functional vigilance
Detect and help manage driver functional vigilance
Driver monitors automation
Vehicle Automation
Driver
Driver chooses role
(based on context)
Automation performs task
Driver performs task
Driving Task
Inform human of elevated probability of automation error
Secondary Supervising System
Estimate probability of automation error
Driver monitors machine
Vehicle Automation
Driver
Driver chooses role
(based on context)
Automation performs task
Driver performs task
Driving Task

(a) Expansion of the functional vigilance framework in the driving context to include driver monitoring and coaching that both detects and manages the functional vigilance of the driver.

(b) Expansion of the Functional vigilance framework in the driving context to include a secondary perception-control system that anticipates automation failure via a mechanism such as the "arguing machines" framework described in [38].

Fig. 6: Two expansions of the functional vigilance framework that may have a positive effect on staving off driver complacency with increasing frequency and duration of automation use.

Tricky Situation	Act Before Tricky Situation	Act After Tricky Situation	All Tricky Situations
Curve	**33.3%**	6.9%	**31.2%**
Cars, pedestrians, objects	**22.7%**	**24.5%**	**22.9%**
Lane merge, split, or shift	**17.9%**	8.2%	**17.2%**
Too close to lane, wall, or car	11.8%	**30.5%**	13.3%
No lane markings	10.7%	4.3%	10.2%
Drifting onto or over lane line	1.2%	**24.5%**	3.0%
Other discomfort or unknown	2.4%	0.7%	2.2%
Object or activity inside car	0.02%	0.4%	0.05%

TABLE II: Categories of tricky situations and their frequencies for preemptive and reactive disengagements. Yellow cells show the top 3 most frequent categories for each disengagement type.

sion of this work is that owners of Tesla vehicles are especially tech-savvy and appear to not include some of the higher-risk demographics such as teenage drivers. This is a relevant point. However, the literature on human behavior in relation to automation [13, 15, 16, 17, 18] observes many of the same patterns across all populations. Therefore, if our findings are not fully generalizable, they are nevertheless highly surprising and informative for the population in our data.

2) Long Term Effects: For drivers to maintain functional vigilance, they have to regulate the degree of their trust in the system such that they don't over-rely on it to a degree that the system cannot functionally support. In our analysis, we present evidence that drivers in our dataset appear to successfully self-regulate Autopilot use and do not appear to over-trust the system in a way that compromises functional vigilance. We have been tracking most of the drivers in our study for over a year and some for over 2 years [12]. However, we do not capture long-term effects that may span 2 or more years. It's possible that as drivers do less and less manual driving, their ability to perform the driving task degrades and their ability to self-regulate an effective collaboration with automation degrades as well. Also, if system reliability were to improve significantly over time, the levels of functional vigilance observed here may well change.

3) Measuring Functional Vigilance: Our approach to measuring functional vigilance focuses on critical events connected to disengagement of Autopilot. As discussed in §III-E, there may be critical event during Autopilot use that do not lead to a disengagement or crash, such as running a red light or drifting lanes (see CE4 in §III-E). It is possible that such events may exist in the dataset and have not yet been detected by our methods. However, they are hypothesized to be rare as compared to the set of tricky situations that were annotated. Conceptually one could perform a rigorous computer vision analysis of this data in order to assess the frequency of these events but any such approach would have its own limitations. Nevertheless, such an effort is something that remains to be done in the future. It is plausible that, as many worry, driver inattention would couple with failures of this type to shape risk but this effect is beyond the focus of this work and remains to be examined in future efforts.

4) Functional Vigilance Decrement and Safety: The results described in this work show that drivers in our dataset maintained sufficient functional vigilance to respond to tricky situations that arose at the time of disengagement. These results do not, however, make any comparative claims on vehicle safety. The fact that drivers in our dataset maintained functional vigilance may or may not be correlated to measures of vehicle safety, but such analysis is outside the scope of this work.

It is also recognized that we are talking about behavior observed in this substantive but still limited naturalistic sample. This does not ignore the likelihood that there are some individuals in the population as a whole who may over-trust a technology or otherwise become complacent about monitoring system behavior no matter the functional design characteristics of the system. The minority of drivers who use the system incorrectly may be large enough to significantly offset the functional vigilance characteristics of the majority of the drivers when considered statistically at the fleet level. The limited number of, but well reported, Tesla crash events make the case that not all drivers develop a sufficient understanding of the realistic ODD characteristics or other limitations of the system, or of the need to maintain an appropriate level of overall situational monitoring. At the same time, it seems apparent that not all drivers of manually controlled vehicles fully appreciate the limitations of such vehicles' capabilities nor exercise appropriate functional vigilance at all times. Crash prevalence on our highways make this clear.

5) Subjective Annotation of Tricky Situations: As described in §III-F, the subjective annotation of epoch aims to label the Autopilot disengagement scenarios that are deemed "tricky" by human observers. This does not necessarily mean that if the driver did not elect to take control when they did that the automation would fail to handle the situation. In other words, the subjective annotation of tricky situations is strictly an approximation of what is a difficult scenario for the automation.

Two guiding directives were given to the annotators for labeling responses as "too late". The primary directive was *Situations where the driver both should have and could have responded in a more timely manner in order to avoid safety-critical consequences of a tricky situation.* The secondary directive was *Any response delayed by more than 1 second after the tricky situation should likely be annotated as "too late."* The exact wording of the directives, emphasis on the first directive, and the choice of 1 second for the second directive all may have had an impact on the annotation result. Future work may include exploration of sensitivity of the final annotation result to the choice and structure of both qualitative and quantitative directives.

C. Path Forward for AI-assisted Vehicle Systems

The findings in this work indicate that functional vigilance does not appear to be decremented during Autopilot use in the dataset under consideration. However, they do not make clear how AI-assisted vehicle systems should be designed to optimize for functional vigilance. We propose two high-level design principles of allowing for (1) exploration and (2) imperfection in §V-A above. Exactly how to implement such principles is an open question. These two principles are likely to be only a subset of what is needed to design safe and enjoyable AI-assisted driving experiences for the entire population of drivers.

We also propose two other potentially highly beneficial additions to aid in the management of the functional vigilance framework as illustrated in Fig. 6. The first is in the form of a feedback loop that includes sensing and managing the state of the driver. This allows the machine to supervise the supervisor and warn them when a functional vigilance decrement, attentional failure, or other deviation from reasonable driver activity is detected. Such supervision can take the form of a hands-on steering wheel sensor and a camera-based driver monitoring system (e.g., Cadillac Super Cruise system). The degree of fidelity and approach most helpful to effectively sense and manage the state of the driver is an important open area of research.

The second proposed addition of potential benefit is a secondary perception-control system as detailed in [38] where a third-party system serves as a supervisor of the primary automation providing the driver with an additional signal on the uncertainty (probability of error) in the sequence of perception and control decisions made by the primary system. Taken together, these additions may help manage the performance of the human (Fig. 6a) and the machine (Fig. 6b) in functional vigilance framework. Furthermore, they may help provide additional protection in the event that there are CE4 events that do exist, but have not yet been identified.

VI. Conclusion

In this work, we provide evidence from a large-scale Tesla Autopilot driving dataset that drivers in this dataset maintain functional vigilance in their use of Autopilot. In this data, drivers use the system extensively and travel 34.8% of miles under Autopilot control. We annotated 18,928 epochs of Autopilot disengagements for presence of challenging scenarios (termed "tricky situations") and whether the driver anticipates or responds to these situations in a timely manner. Through this process, 8,729 tricky situations were annotated by multiple individuals until an agreement was reached about the timeliness of the driver response with respect to the critical signal. The resulting annotation categorized the driver as functionally vigilant in all cases as determined through subjective annotation of the temporal characteristics of their response. We discuss two possible contributing factors that underlie these results, and propose possibly beneficial expansion of the functional vigilance framework with monitoring systems for both the human and the automation. Finally, we enumerate several limitations of our work including that these findings (1) cannot be directly used to infer safety as a much larger dataset would be required for crash-based statistical analysis of risk, (2) may not be generalizable to a population of drivers nor Autopilot versions outside our dataset, (3) do not include challenging scenarios that did not lead to Autopilot disengagement, (4) are based on human-annotation of critical signals, and (5) do not imply that driver attention management systems are not potentially highly beneficial additions to the functional vigilance framework for the purpose of encouraging the driver to remain appropriately attentive to the road.

The authors are highly cognizant that there are significant nuances in the design, analysis, and interpretation of this work. It is our hope that it will encourage serious discussion and further investigation of how seemingly subtle features of AI-assisted system design and implementation may influence the extent to which humans are able to sustain appropriate collaborative engagement with such technology.

Future work will include automated glance region classification of drivers during Autopilot control and comparable baseline periods during manual control in order to gain a greater depth of insight on attention allocation and functional vigilance decrement during prolonged use of Autopilot. In addition, extensive self-report data collection through questionnaires of Tesla owner will be conducted to gain an understanding of system use and perceptions across a large population of drivers.

VII. Acknowledgement

Support for this work was provided by the Advanced Vehicle Technology (AVT) consortium at MIT. The views and conclusions being expressed are those of the authors and may not necessarily represent those of individual sponsoring organizations. All authors listed as affiliated with MIT contributed to the work only during their time at MIT as employees or visiting graduate students.

The authors would like to thank Charles Green, Jeffrey Blecher, and many anonymous reviewers for helpful feedback and suggestions through the many iterations of the work. In addition, we would like to thank the entire team at MIT behind the development and support of the Autonomous Vehicle Technology (MIT-AVT) Study [12].

References

[1] S. international, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," *SAE International,(J3016)*, 2016.

[2] A. Gray, Y. Gao, J. K. Hedrick, and F. Borrelli, "Robust predictive control for semi-autonomous vehicles with an uncertain driver model," in *2013 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2013, pp. 208–213.

[3] M. Blanco, J. Atwood, H. M. Vasquez, T. E. Trimble, V. L. Fitchett, J. Radlbeck, G. M. Fitch, S. M. Russell, C. A. Green, B. Cullinane *et al.*, "Human factors evaluation of level 2 and level 3 automated driving concepts," Tech. Rep., 2015.

[4] C. Gold, D. Damböck, K. Bengler, and L. Lorenz, "Partially automated driving as a fallback level of high automation," in *6. Tagung Fahrerassistenzsysteme*, 2013.

[5] "Euro NCAP 2018 automated driving tests," http://bit.ly/2Wxinhp, accessed: 2019-03-30.

[6] A. Roy, "The language of self-driving cars is dangerous – here's how to fix it," http://bit.ly/2WF86Qx, accessed: 2019-03-30.

[7] H. Abraham, B. Seppelt, B. Mehler, and B. Reimer, "What's in a name: Vehicle technology branding & consumer expectations for automation," in *Proceedings of the 9th international conference on automotive user interfaces and interactive vehicular applications*. ACM, 2017, pp. 226–234.

[8] L. Fridman, "Tesla Vehicle Deliveries and Autopilot Mileage Statistics," Jan. 2019.

[9] M. Dikmen and C. M. Burns, "Autonomous driving in the real world: Experiences with tesla autopilot and summon," in *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 2016, pp. 225–228.

[10] M. R. Endsley, "Autonomous driving systems: A preliminary naturalistic study of the tesla model s," *Journal of Cognitive Engineering and Decision Making*, vol. 11, no. 3, pp. 225–238, 2017.

[11] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 182–193, 2016.

[12] L. Fridman, D. E. Brown, M. Glazer, W. Angell, S. Dodd, B. Jenik, J. Terwilliger, J. Kindelsberger, L. Ding, S. Seaman, A. Mehler, A. Sipperley, A. Pettinato, B. Seppelt, L. Angell, B. Mehler, and B. Reimer, "MIT autonomous vehicle technology study: Large-scale deep learning based analysis of driver behavior and interaction with automation," *CoRR*, vol. abs/1711.06976, 2017. [Online]. Available: http://arxiv.org/abs/1711.06976

[13] N. H. Mackworth, "The breakdown of vigilance during prolonged visual search," *Quarterly Journal of Experimental Psychology*, vol. 1, no. 1, pp. 6–21, 1948.

[14] K. R. Boff, L. Kaufman, and J. P. Thomas, "Handbook of perception and human performance. volume 2. cognitive processes and performance," Harry G Armstrong Aerospace Medical Research Lab, Tech. Rep., 1994.

[15] D. W. Bates, M. Cohen, L. L. Leape, J. M. Overhage, M. M. Shabot, and T. Sheridan, "Reducing the frequency of errors in medicine using information technology," *Journal of the American Medical Informatics Association*, vol. 8, no. 4, pp. 299–308, 2001.

[16] L. R. Hartley, P. Arnold, H. Kobryn, and C. MacLeod, "Vigilance, visual search and attention in an agricultural task," *Applied ergonomics*, vol. 20, no. 1, pp. 9–16, 1989.

[17] M. R. Endsley and E. O. Kiris, "The out-of-the-loop performance problem and level of control in automation," *Human factors*, vol. 37, no. 2, pp. 381–394, 1995.

[18] R. Molloy and R. Parasuraman, "Monitoring an automated system for a single failure: Vigilance and task complexity effects," *Human Factors*, vol. 38, no. 2, pp. 311–322, 1996.

[19] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human factors*, vol. 52, no. 3, pp. 381–410, 2010.

[20] R. Parasuraman, R. Molloy, and I. L. Singh, "Performance consequences of automation-induced "complacency"," *The International Journal of Aviation Psychology*, vol. 3, no. 1, pp. 1–23, 1993.

[21] N. Bagheri, G. A. Jamieson *et al.*, "Considering subjective trust and monitoring behavior in assessing automation-induced "complacency."," *Human performance, situation awareness, and automation: Current research and trends*, pp. 54–59, 2004.

[22] C. D. Wickens and S. R. Dixon, "The benefits of imperfect diagnostic automation: A synthesis of the literature," *Theoretical Issues in Ergonomics Science*, vol. 8, no. 3, pp. 201–212, 2007.

[23] P. A. May, "Effects of automation reliability and failure rate on monitoring performance in a multi-task environment," Ph.D. dissertation, Catholic University of America, 1993.

[24] M. W. Wiggins, "Vigilance decrement during a simulated general aviation flight," *Applied Cognitive Psychology*, vol. 25, no. 2, pp. 229–235, 2011.

[25] J. A. Caldwell, "Fatigue in aviation," *Travel medicine and infectious disease*, vol. 3, no. 2, pp. 85–96, 2005.

[26] R. L. Helmreich, "On error management: lessons from aviation," *Bmj*, vol. 320, no. 7237, pp. 781–785, 2000.

[27] B. Reimer, A. Pettinato, L. Fridman, J. Lee, B. Mehler, B. Seppelt, J. Park, and K. Iagnemma, "Behavioral impact of drivers' roles in automated driving," in *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 2016, pp. 217–224.

[28] Z. Lu, R. Happee, C. D. Cabrall, M. Kyriakidis, and J. C. de Winter, "Human factors of transitions in automated driving: A general framework and literature survey," *Transportation research part F: traffic psychology and behaviour*, vol. 43, pp. 183–198, 2016.

[29] E. T. Greenlee, P. R. DeLucia, and D. C. Newton, "Driver vigilance in automated vehicles: hazard detection failures are a matter of time," *Human factors*, vol. 60, no. 4, pp. 465–476, 2018.

[30] O. Carsten, F. C. Lai, Y. Barnard, A. H. Jamson, and N. Merat, "Control task substitution in semiautomated driving: Does it matter what aspects are automated?" *Human factors*, vol. 54, no. 5, pp. 747–761, 2012.

[31] I. S. Marcos, *Challenges in Partially Automated Driving: A Human Factors Perspective*. Linköping University Electronic Press, 2018, vol. 741.

[32] A. af Wåhlberg, *Driver behaviour and accident research methodology: unresolved problems*. CRC Press, 2017.

[33] J. D. Lee, K. L. Young, and M. A. Regan, "Defining driver distraction," *Driver distraction: Theory, effects, and mitigation*, vol. 13, no. 4, pp. 31–40, 2008.

[34] F. Guo, S. G. Klauer, M. T. McGill, and T. A. Dingus, "Evaluating the relationship between near-crashes and crashes: Can near-crashes serve as a surrogate safety metric for crashes?" 2010.

[35] R. E. Llaneras, B. R. Cannon, and C. A. Green, "Strategies to assist drivers in remaining attentive while under partially automated driving: Verification of human–machine interface concepts," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2663, pp. 20–26, 2017.

[36] T. W. Victor, E. Tivesten, P. Gustavsson, J. Johansson, F. Sangberg, and M. Ljung Aust, "Automation expectation mismatch: incorrect prediction despite eyes on threat and hands on wheel," *Human factors*, vol. 60, no. 8, pp. 1095–1116, 2018.

[37] F. H. Administration, "Highway functional classification concepts, criteria and procedures," 2013.

[38] L. Fridman, L. Ding, B. Jenik, and B. Reimer, "Arguing machines: Human supervision of black box ai systems that make life-critical decisions," *CoRR*, vol. abs/1710.04459, 2019. [Online]. Available: http://arxiv.org/abs/1710.04459

[39] M. Buehler, K. Iagnemma, and S. Singh, *The DARPA urban challenge: autonomous vehicles in city traffic*. springer, 2009, vol. 56.