2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks

Hands Off the Wheel in Autonomous Vehicles?

A Systems Perspective on over a Million Miles of Field Data

Subho S. Banerjee[§], Saurabh Jha[§], James Cyriac[†], Zbigniew T. Kalbarczyk[†] and Ravishankar K. Iyer[†§]
[§]Department of Computer Science, [†]Department of Electrical and Computer Engineering,
University of Illinois at Urbana-Champaign, Urbana IL - 61801, USA.

***Abstract*—Autonomous vehicle (AV) technology is rapidly becoming a reality on U.S. roads, offering the promise of improvements in traffic management, safety, and the comfort and efficiency of vehicular travel. The California Department of Motor Vehicles (DMV) reports that between 2014 and 2017, manufacturers tested 144 AVs, driving a cumulative 1,116,605 autonomous miles, and reported 5,328 disengagements and 42 accidents involving AVs on public roads. This paper investigates the causes, dynamics, and impacts of such AV failures by analyzing disengagement and accident reports obtained from public DMV databases. We draw several conclusions. For example, we find that autonomous vehicles are 15 – 4000× worse than human drivers for accidents per cumulative mile driven; that drivers of AVs need to be as alert as drivers of non-AVs; and that the AVs' machine-learning-based systems for perception and decision-and-control are the primary cause of 64% of all disengagements.**

***Index Terms*—Autonomous Vehicles, Reliability, Fault Characterization, Disengagement, Accident.**

I. Introduction

Autonomous vehicle (AV) technologies are advertised to be transformative, with a potential to improve traffic congestion, safety, productivity, and comfort [1]. Several states in the U.S. (e.g., California, Texas, Nevada, Pennsylvania, and Florida) have already started testing AVs on public roads. Prior research into AVs has focused predominantly on the design of automation technology [2]–[7], its adoption [8], the impact of AVs on congestion [9], and the legal [10], [11] and regulatory barriers [12]–[15] for AV implementation. With the increasing popularity and ubiquitous deployment of semi- and fully-automated vehicles on public roads, safety and reliability have increasingly become critical requirements for public acceptance and adoption. This paper assesses, in broad terms, the reliability of AVs by evaluating the cause, dynamics, and impact of failures across a wide range of AV manufacturers utilizing publicly available field data from tests on California public roads, including urban streets, freeways, and highways.

Dataset. The California Department of Motor Vehicles (CA DMV) mandates that all manufacturers testing AVs on public roads file annual reports detailing *disengagements* (a failure that causes the control of the vehicle to switch from the software to the human driver) and *accidents* (an actual collision with other vehicles, pedestrians, or property) [16]. The focus of the testing program, and of this paper, is on semi-autonomous vehicles that require a human driver to serve as a fall-back in the case of failure. In particular, we are interested in studying failures that pertain to sensing (e.g., cameras, LIDAR) and computing systems (e.g., hardware and software systems that enable environment perception and vehicle control) that enable the "self-driving" features of the vehicles. We analyze field data collected over a 26-month period from September 2014 to November 2016 (part of the DMV's 2016 and 2017 data releases), containing data from 12 AV manufacturers for 144 vehicles that drove a cumulative $1,116,605$ autonomous miles. Across all manufacturers, we observe a total of $5,328$ disengagements, 42 of which led to accidents.

Results. This paper presents 1) an end-to-end workflow for analyzing AV failure data, and 2) several insights about failure modes in AVs (across a single manufacturer's fleet, across different manufacturers, and in time) by executing the proposed workflow on the available data. Our study shows:

1) Drivers of AVs need to be as alert as drivers of non-AV vehicles. Further, the small size of the overall action window (detection time + reaction time) would make reaction-time-based accidents a frequent failure mode with the widespread deployment of AVs.
2) For the same number of miles driven, for the manufacturers that reported accidents, human-driven non-AVs were $15 - 4000\times$ less likely than AV's to have an accident.
3) 64% of disengagements were the result of problems in, or untimely decisions made by, the machine learning system.
4) In terms of reliability per mission, AVs are $4.22\times$ worse than airplanes, and $2.5\times$ better than surgical robots.

These findings demonstrate that while individual components of AV technology (e.g., vision systems, control systems) may have matured, entire AV systems are still in a "burn-in" phase.

The analysis presented in this paper shows a distinct improvement in the performance of AVs over time. However, it also demonstrates the need for continued improvement in the dependability of this technology. It is conceivable (moreover, expected) that AV manufacturers are performing a similar analysis of data coming from their testing fleets, but to the best of our knowledge, information on such analysis is not available publicly. Our goal is to support resilience research by characterizing failures of autonomous vehicles, rather than to further the operational perspective of the manufacturer. Our results can better inform the design of future AVs.

Organization. Fig. 1 shows the end-to-end pipeline for processing failure data from autonomous vehicles. Section II describes two real examples of AV-related accidents on California roads. Section III describes the AVs and the data collection methodology (*Stage I* of the pipeline). Section IV describes the preprocessing, filtering, and natural language processing (NLP) steps required to convert the data to a format suitable for analysis (*Stages II & III* of the pipeline). Section V describes the statistical analysis of the failure data and summarizes the insights derived from the analysis (*Stage IV* of the pipeline). Finally, Sections VI to VIII describe the threats to validity, related work and conclusions, respectively.

2158-3927/18/$31.00 Â©2018 IEEE
DOI 10.1109/DSN.2018.00066

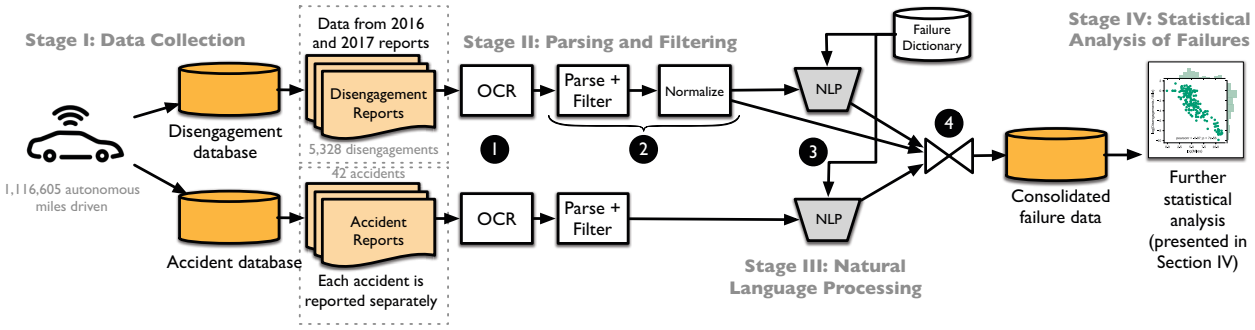


Figure 1. The end-to-end data collection, processing, and analysis pipeline that forms the basis of this study.

II. Case Studies

In this section, we present two representative case studies based on real events that occurred in the streets of Mountain View, CA. These case studies illustrate how problems in the perception, learning, and control systems of an AV can manifest as an accident.

A. Case Study I: Real-Time Decisions

Example 1 in Fig. 2 shows a case in which the human driver of the AV proactively took over the control of the vehicle from the autonomous agent (to prevent an accident) but was unable to rectify decisions made by the autonomous agent in time to prevent an accident. The disengagement report (i.e., error logs from the AV combined with post-mortem analysis performed by the manufacturer) logs the error as either "`Disengage for a recklessly behaving road user`" or "`incorrect behavior prediction`." Specifically, a Waymo prototype vehicle was in autonomous mode at a street intersection when a pedestrian started to cross the street. From the accident report, we find that the AV decided to yield to the pedestrian but did not stop. The test driver proactively took control of the car as a precaution. At the same time, there was a car in front of the AV that was also yielding to the pedestrian, and another vehicle to the rear in the adjacent lane that was making a lane change. In this complex scenario, the driver did not have many options other than to brake, and the rear vehicle collided with the back of the AV.

B. Case Study II: Anticipating AV Behavior

Example 2 in Fig. 2 shows a case in which a Waymo prototype vehicle was running in autonomous mode and was hit by a manual vehicle from the rear at a street intersection. The disengagement report logs the cause as "`Disengage for a recklessly behaving road user`." In this case, the AV had signaled a right turn and had started to decelerate for the turn. It came to a complete stop before it started moving again towards the intersection to gauge the traffic coming from the other side in order to make a safe turn. The movement towards the intersection was required to allow the recognition system to analyze the scene and produce a movement plan for the car. The driver of the rear vehicle was confused and interpreted this movement to mean that the AV was conitinuing on its path (i.e., making the turn). The driver first stopped (as the AV stopped) and then started moving (as the AV started to move again). This resulted in a rear collision on the AV, as the driver could not anticipate the actions of the AV.

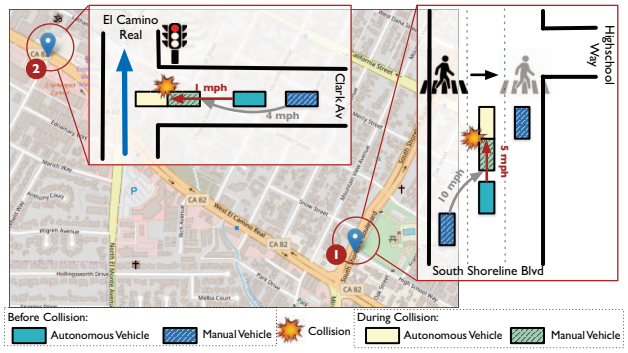


Figure 2. Accident scenarios.

C. Summary

By law, both of those accidents were caused by the drivers in the non-AV; however, close inspection of the accident reports shows that the AV had a significant share of the responsibility. The above examples showcase the poor AV decision-making that eventually leads to accidents.

1) The street intersections represent complex scenarios in which the AV needs to analyze multiple traffic flows and make decisions in a constrained environment. Based on our analysis we attribute the failures to the learning-based perception system, which did not infer in time the evolving environment dynamics from the onboard sensor systems (e.g., RADAR, LIDAR), leading the learning-based control system to make inadequate decisions.
2) In both cases, drivers either voluntarily took or were forced to take control from the autonomous system in complex and dynamic traffic scenarios that frequently gives them very little time to react and undo the AV's actions. The perception and reaction time is crucial in accident avoidance.
3) Drivers in other non-AVs often cannot anticipate decisions made by AVs, which frequently also leads to accidents.

Using the limited publicly available information about the design of the AV systems (e.g., [17]–[20]), we draw our conclusions by analyzing human-entered textual logs that contain information about accidents and disengagements. Our method localizes failures to the learning, perception, and decision-and-control subsystems of an AV to understand the causes of disengagements and accidents.

III. AV System Description and Data Collection

A. Preliminaries

1) Autonomous Vehicles

An AV is any vehicle that uses an autonomous driving system (ADS) technology capable of supporting and assisting a human driver in the tasks of 1) controlling[1] the main functions of steering and acceleration, and 2) monitoring the surrounding environment (e.g., other vehicles/pedestrians, traffic signals, and road markings) [21].

The Society of Automotive Engineers (SAE) defines six levels of autonomy that are based on the extent to which the technology is capable of supporting and assisting the driving tasks [21]. The levels of autonomy go from 0 (no automation) to 5 (full, unrestricted automation). Levels 0–2 (e.g., anti-lock braking, cruise control) require a human driver to be responsible for monitoring the environment of the vehicle, with different levels of automation available to support vehicle control tasks. Levels 3–5 are thought of as truly automated driving systems where the AV both monitors the environment and controls the vehicle. The subject of this paper is the Level 3 vehicles.

2) Disengagements

Level 3 requires the presence (and attention) of a human driver to serve as a fall-back when the autonomous system fails. A transfer of control from the autonomous system to the human driver in the case of a failure is called a *disengagement*. Disengagements can be initiated either manually by the driver or autonomously by the car. Manual disengagements initiated by the driver are cautionary (e.g., if one feels uncomfortable, or wants to adopt a proactive approach to prevent a potential accident). Automated disengagements are indicative of a design limitation of the AV.

3) Accidents

An *accident* is an actual collision with other vehicles, pedestrians, or property. Note that not all disengagements lead to collisions. As we show later in this paper, most disengagements are handled safely by the human operators, with only a small fraction leading to accidents. For example, in some reported collisions, the test driver initiated a manual disengagement before the collision (an artifact of the training program that all test drivers acting as AV safety-pilots have to undergo before they are allowed on public roads [16]).

B. AV Hierarchical Control Structure

Manufacturers have not disclosed the architectures of their autonomous vehicles. However, to identify multidimensional causes of AV disengagements/accidents, we built a hierarchical control structure for AVs by using the systems-theoretic hazard modeling and analysis abstraction STPA (Systems-Theoretic Process Analysis) [23]. Fig. 3 shows an AV hierarchical control structure derived based on technical documentation [22], [24]–[27]. We assert that these information sources are representative and provide a conceptual view of AV systems that is sufficiently detailed to enable creation of an STPA model. We refer to this system as the "Autonomous Driving System" (ADS). The major components of the ADS are 1) "sensors" (e.g., GPS, RADAR, LIDAR, and cameras) that are responsible for collecting environment-related data, 2) a "recognition system"[2] that uses sensor data to identify the objects and changes in the environment around the AV, 3) a "planner and controller" system that is responsible for planning the next motion of the car based on the current parameters of the AV and the environment (e.g., speed, location, and other vehicles), and 4) a "follower" system that signals the "actuators" to drive the vehicle along the path chosen by the "planner and controller."

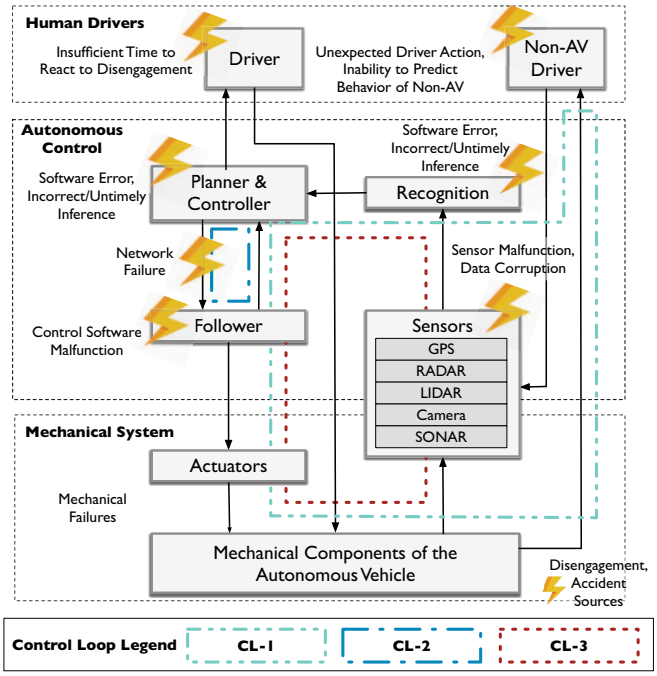


Figure 3. Autonomous vehicle hierarchical control structure drawn based on [22]. Examples of control loops are highlighted as `CL-1`, `CL-2`, and `CL-3`.

STPA employs concepts from systems and control theories to model hierarchical control structures in which the components at each level of the hierarchy impose safety constraints on the activity of the levels below and communicate their conditions and behavior to the levels above them. Accidents and disengagements are complex dynamic processes resulting from inadequate perception control and decision-making at different layers of the system control structure. Accidents and disengagements seen in the data were overlaid on this structure.

In every control loop, the planner and controller system uses an algorithm to generate the control actions based on a model of the current state of the process that it is controlling. The control actions (e.g., "decelerate") taken by the planner and controller system (i.e., the autonomous driving system) change the state of the controlled process (e.g., mechanical components of the autonomous vehicle). The feedback message (e.g., the state of the traffic lights) sent back from the controlled process (e.g., the AV control software) updates the process model used (e.g., the mental model the driver has of the AV status) by the controller. Analysis of dependencies along those control loops allows for the identification of inadequate controls and the potential causes of those unsafe control actions through examination of the operation of components and their interactions in each

[1]Here, "control" incorporates both decision and control.

[2]The "recognition system" is also referred to as the "perception system."

Table I
SUMMARIZATION OF FLEET SIZE, AUTONOMOUS MILES DRIVEN, AND FAILURE INCIDENTS ACROSS ALL MANUFACTURERS IN THE DATASET.

	2015–2016 Report				2016–2017 Report			
Manufacturer	**Cars**	**Miles**	**Disengagments**	**Accidents**	**Cars**	**Miles**	**Disengagements**	**Accidents**
Mercedes-Benz	2	1739.08	1024	–	–	673.41	336	–
Bosch	2	935.1	625	–	3	983	1442	–
Delphi	2	16661	405	1	2	3090	167	–
GM Cruise	–	285.4	135	–	–	9729.8	149	14
Nissan	4	1485.4	106	–	3	4099	29	1
Tesla	–	–	–	–	5	550	182	–
Volkswagen	2	14946.11	260	–	–	–	–	–
Waymo (Google)	49	424332	341	9	70	635868	123	16
Uber ATC	–	–	–	–	–	–	–	1
Honda	–	–	–	–	0	0	0	–
Ford	–	–	–	–	2	590	3	–
BMW	–	–	–	–	–	638	1	–
Total	61	460384.1	2896	10	83	656221	2432	32

Dashes indicate the absence of data in the manufacturer's report.

loop of the control structure. Any flaws or inadequacies in the algorithm, the process model, or the feedback used by a controller are considered potential causal factors leading to unsafe control actions and resultant disengagements/accidents.

In Fig. 3 we highlight three control loops (`CL-1`, `CL-2`, and `CL-3`, indicated with different types of dashed lines) to illustrate details of the interactions among the driver (both AV and Non-AV), AV control, and AV hardware/software components. Our analysis couples that STPA approach with manufacturers' reports. The most complex control loop, `CL-1`, involves interaction among the *autonomous control* (including sensors, recognition system, planner, and controller), *mechanical system* (actuators and mechanical components of the vehicle), and *human drivers* (drivers of non-AVs). The Non-AV Driver module represents the AV system's ability to 1) collect the data on Non-AV driver behavior through the sensors, and 2) provide information (e.g., on brake signals, turn indicators, or horn) to Non-AV drivers. Examples of failures in this control loop were discussed in the two case studies presented earlier.

C. Data Sources

The CA DMV is the state agency that registers motor vehicles, issues regulations and permits, and monitors the testing and field operation of autonomous vehicles. California driving conditions are representative of urban situations and the DMV has a strong mandate for data collection and public availability. California law requires the manufacturers operating and testing AVs to file reports on disengagements (reported annually) and accidents (reported within ten business days of the incident) [16], [28]; these reports are eventually made public. The reports are available as a part of two databases:

1) *AV Disengagement Reports*: These reports contain aggregated information about fleet size, monthly autonomous miles traveled, and the number of disengagements observed. Each manufacturer provides its own data format, resulting in a fragmented set of data. Some manufacturers provide additional information, including timestamps, road type (e.g., urban streets, highway, freeway), weather conditions (e.g., sunny, raining, overcast), driver reaction times (time taken for the driver to disengage from autonomous mode), and other factors contributing to the disengagements. We use the additional data whenever it is available.
2) *AV Accident Reports*: These reports contain timestamped information about the autonomous vehicle involved, the location of the accident, descriptions of other vehicles involved (e.g., class of vehicle, speed), and human-written textual description of the incident and its severity.

Both datasets consist of scanned documents containing both tabulated data and natural-language text. Unlike previous analyses [29], [30], which are based solely on the data provided, we focus on building an analysis workflow that processes substantive amounts of human-generated disengagement and accident reports by using NLP.

Summary of Datasets. The datasets cover 12 AV manufacturers (Bosch, Delphi Automotive, Google, Nissan, Mercedes-Benz, Tesla Motors, BMW, GM, Ford, Honda, Uber, and Volkswagen). With 144 AVs that drove a cumulative $1,116,605$ autonomous miles across 9 distinct road types (31.7% on city streets, 29.26% on highways, 14.63% on interstates, 9.75% on freeways, and the remaining 14.6% in parking lots, suburban, and rural roads). Uber, BMW, Ford, and Honda reported too few disengagements for us to draw statistically significant conclusions, so are left out of the analysis in this paper. Across all manufacturers, we observe a total of $5,328$ disengagements[3] and 42 accidents (including the two case studies in Section II). Aggregating per car and per manufacturer, we observe an average of 262 autonomous miles driven per disengagement, and one accident event for every 127 disengagements.

Across manufacturers in the dataset, we observe a significant skew in the number of autonomous miles driven (see Table I). For example, Waymo tested their AV prototypes more extensively than the others (over 1,000,000 miles compared to 15,000 miles for the next highest testing manufacturer). This suggests that Waymo's AVs might perform better than those of its competitors because of the extensive testing of the ADS platform. Note that not all manufacturers provide all the data needed to compute the summary statistics; those omissions are indicated by dashes in Table I.

[3]Two of the manufacturers (Bosch and GMCruise) reported all their disengagement data as planned tests. Our understanding, based on all the DMV reports, is that the tests were planned, but the disengagements occurred naturally. Together the two manufacturers have 14 accidents during "tests".

Table II
SAMPLE OF DISENGAGEMENT REPORTS FROM THE CA DMV DATASET.

Manufacturer	Raw Disengagement Report (Log)	Category	Tags
Nissan	1/4/16 — 1:25 PM — **Software module froze**. As a result driver safely disengaged and resumed manual control. — City and highway — Sunny/Dry	System	Software
Nissan	5/25/16 — 11:20 AM — Leaf #1 (Alfa) — The AV **didn't see** the lead vehicle, driver safely disengaged and resumed manual control.	ML/Design	Recognition System
Waymo	May-16 — Highway — Safe Operation — Disengage for a **recklessly behaving** road user	ML/Design	Environment
Volkswagen	11/12/14 — 18:24:03 — Takeover-Request — **watchdog error**	System	Computer System

We use the "—" to denote field separators.
Note that log formats vary across manufacturers and time.
Bold-face text represents phrases analyzed by the NLP engine to categorize log lines.

IV. DATA-ANALYSIS WORKFLOW: PARSING, FILTERING, NORMALIZATION AND NLP

Fig. 1 describes our methodology (workflow) for converting raw disengagement and accident reports into a consolidated form that lends itself to further analysis. Below, we describe the key steps involved in Stages II and III of the workflow.

Digitization of the Accident and Disengagement Reports. The aforementioned logs are provided in the form of scanned images of digital documents (for disengagement reports) and hand written reports (for accident reports). The first task is to pre process and convert these scanned reports into a machine-encoded format. Examples of such machine-encoded disengagement reports are shown in Table II. Hence, our analysis proceeds with optical character recognition (OCR; labeled as ❶ in Fig. 1) by using Google Tesseract [31] on the scanned documents. In certain cases, where the Teserract OCR failed (because of low-resolution scans or inability to recognize some table formats), we manually converted the documents to machine-encoded text.

Data Normalization. CA DMV regulations require that each manufacturer report crucial information about disengagements, e.g., the number of miles driven in autonomous mode and the number of disengagements observed. However, it does not enforce any data format specification for these reports, leading to disparities (across manufacturers and across time) in the data schema and granularity of the information available through these reports. Hence, we need to filter, parse, and normalize (labeled as ❷ in Fig. 1) the data into machine-encoded text to produce structured datasets that have uniform schema across manufacturers and time (i.e., across reports made by the same manufacturer at different times). Taken together, steps ❶ and ❷ correspond to preprocessing of the datasets to make them ready for further analysis.

Labeling and Tagging of the Reported Disengagement and Accident Causes. The pipeline uses an NLP-based technique (labeled as ❸ in Fig. 1) to map a given disengagement event in a corresponding fault tag and a failure category. First we make several passes over the dataset to construct a "Failure Dictionary" that contains a sequence of phrases (keywords) extracted from the raw disengagement reports (logs). This dictionary is used to design a voting scheme (which is based on the maximum number of shared keywords) to assign a disengagement cause to a fault tag. In the event that this procedure is unsuccessful and we cannot associate any of the known tags to textual description, the disengagement cause is marked with the "Unknown-T" tag.

We then build an ontology (based on Fig. 3) of failure *categories* on top of the tags (which were derived from [32]). Specifically, we apply our understanding of the ADS system (described in Section III-B) to select keywords and phrases that differentiate fault tags from each other. The tags are chosen to localize faults in the computing system (e.g., software and hardware systems) and in the machine learning algorithms/design (e.g., perception and control algorithms), thereby identifying potential targets for improving the safety and reliability of the AV. Table III lists the fault tags used in this study. Table II provides examples of the raw log to tag and category mapping. We consider the following failure categories: 1) faults in the design of the machine learning system responsible for "perception" tasks (dealing with data from sensors) and "planning and control" tasks (dealing with control of steering and acceleration); 2) faults in the computing system (dealing with hardware and software problems); and 3) an "Unknown-C" category consisting of tags we cannot classify into any of the above categories.

These tags and categories allow us to classify the types of fail-

Table III
DEFINITION OF FAULT TAGS AND CATEGORIES THAT ARE ASSIGNED TO DISENGAGEMENTS.

Tag	Category	Definition
Environment	ML/Design	Sudden change in external factors (e.g., construction zones, emergency vehicles, accidents)
Computer System	System	Computer-system-related problem (e.g., processor overload)
Recognition System	ML/Design	Failure to recognize outside environment correctly
Planner	ML/Design	Planner failed to anticipate the other driver's behavior
Sensor	System	Sensor failed to localize in time
Network	System	Data rate too high to be handled by the network
Design Bug	ML/Design	AV was not designed to handle an unforeseen situation
Software	System	Software-related problems such as hang or crash
AV Controller	System	"System" when AV controller does not respond to commands
	ML/Design	"ML/Design" when AV controller makes wrong decisions/predictions
Hang/Crash	System	Watchdog timer error

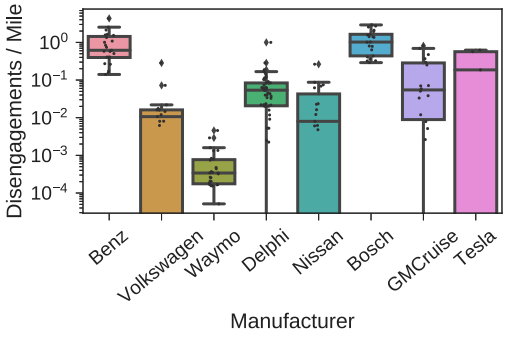


Figure 4. Comparison of the distributions of DPM per car across manufacturers. The boxes show quartiles; the notches show medians; and the whiskers show max/mins.

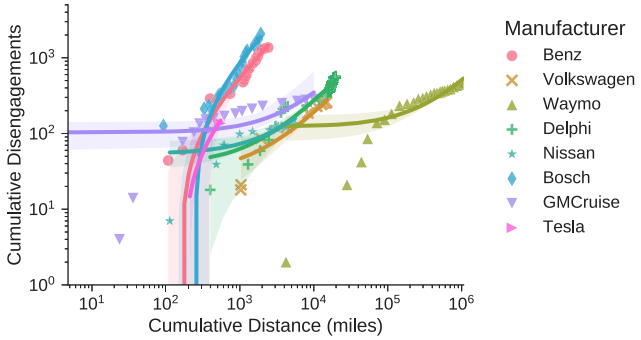


Figure 5. Disengagements reported per cumulative miles driven across manufacturers represented in a log-log plot. Lines represent linear regression fits.

ure causes into machine-learning vs. computer-system-related issues. Table III provides a mapping between the categories and tags used in our analysis. In the final step (labeled as ❹ in Fig. 1), the preprocessed data from the disengagement dataset and accident dataset are merged together, along with extracted categories and tags, to create a *consolidated* AV failure database for further analysis.

V. Statistical Analysis of Failures in AVs

Traditional approaches to evaluating the resilience of a system [33] require the computation of *availability*, *reliability*, and *safety*. These metrics require information about operational periods of the AV (e.g., the active time of the vehicle). As this information is not available in the CA DMV dataset, we use the 5,324 disengagements (across eight manufacturers) and 42 accidents as the basis for deriving statistics on fault classes, failure modes of AVs, and their evolution over time. These statistics allow us to draw conclusions and answer the following questions:

Question 1. How do we assess the stability/maturity of the AV technology?

Question 2. What is the primary cause of disengagements (and potentially accidents) observed in AVs?

Question 3. Are manufacturers indeed building better and more reliable AVs over time?

Question 4. What level of alertness[4] of the human driver of an AV guarantees safety?

Question 5. How well do AVs compare with human drivers?

A. Analysis of AV Disengagement Reports

1) Question 1: Assessment of AV Technology

Based on the available data, we computed the following metrics from the disengagement reports to assess AVs: 1) number of disengagements observed per autonomous mile driven (DPM, shown in Fig. 4), and 2) total number of disengagements observed (shown in Fig. 5).

Comparing DPMs across Manufacturers. Most manufacturers have a median DPM $\in [0.1, 0.01]$ m^{-1} per car with the 99$^{\text{th}}$ percentile DPM around 1 m^{-1} (see Fig. 4). There is a significant disparity (nearly $100\times$) between median DPMs across all manufacturers. This substantiates our initial hypothesis (from Section III-C) that the cumulative miles driven by a manufacturer (see Table I) is indicative of better performance. For example, Waymo (Google) does $\sim 100\times$ better than its competitors in terms of both the median and 99$^{\text{th}}$ percentile DPMs; at the same time, it is responsible for $> 90\%$ of the total miles driven in the dataset.

Maturity of AV Technology. Fig. 5 demonstrates a strong linear correlation (based on the linear regression fits) between the number of disengagements observed and the number of cumulative autonomous miles driven. We expect that in an ideal case mature AV technology will show a decrease in DPM (i.e., the slopes of the lines in Fig. 5) that asymptotically reaches towards a horizontal line (or close to it, i.e., zero DPM or a very low DPM). The reason is that the data collected from the planned testing of AVs validates the computing system (e.g., by identifying software bugs) and also trains the machine learning algorithms that monitor the environment and control the steering and acceleration of the AV. Thereby eventually enabling the AVs to handle more fault scenarios, thus contributing to a decreasing DPM. This is true for most manufacturers to varying degrees with the exception of Volkswagen, Bosch, and GMCruise. *An important conclusion is that despite the million miles driven, Waymo is still not quite approaching the target asymptote. This indicates that Waymo and other manufacturers are still in the "burn-in" phase.*

2) Question 2: Causes of AV Disengagements

We present a categorization of the sources of faults that cause disengagements from two different perspectives: 1) cause of occurrence, and 2) modality of occurrence.

Machine-Learning-Related Faults. First, we consider disengagements by *cause of occurrence*, i.e., categorization of the cause of a disengagement. In the following text, we ignore the numbers for Tesla, as most of their categorical label are marked "Unknown-C." We observe that machine-learning-related faults, mainly ones pertaining to the perception system (e.g., improper detection of traffic lights, lane markings, holes, and bumps), are the dominant cause of disengagements across most manufacturers. They account for $\sim 44\%$ of all reported disengagements (see Table IV).[5] The second major contributor to reported disengagements is the machine learning

[4]Measured here as reaction times of human drivers in case of disengagements.

[5]We consider external fault sources such as undetected construction zones, cyclists, pedestrians, emergency vehicles, and weather phenomena (e.g., rain or sun glare) as perception-related-machine-learning related disengagements as they deal with interpretation of the environment from sensor data.

Table IV
DISENGAGEMENTS ACROSS MANUFACTURERS (AS PERCENTAGES) CATEGORIZED BY ROOT FAILURE CATEGORIES.

Manufacturer	Fault Type			
	ML/Design		System	Unknown-C
	Planner/ Controller	Perception/ Recognition		
Delphi	37.59	50.17	12.24	0
Nissan	36.3	49.63	14.07	0
Tesla	0	0	1.65	98.35
Volkswagen	0	3.08	83.08	13.85
Waymo	10.13	53.45	36.42	0

ML/Design is divided into Planner/Controller- and Perception-related problems.

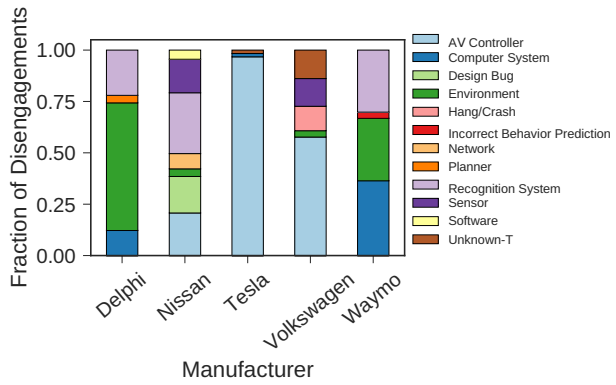


Figure 6. Categorization (in terms of fault tags) of faults that led to disengagements across manufacturers.

related to the control and decision framework (e.g., improper motion planning), which accounts for $\sim 20\%$ of the total disengagements. The computing system, i.e., hardware issues (e.g., problems with the sensor and processor) and software issues (e.g., hangs, crashes, bugs), accounts for $\sim 33.6\%$ of the total disengagements reported. Further, we observe that the perception-based machine learning faults are responsible for DPM measurements in the upper three quartiles. *Therefore we conclude that the faults in the perception system are directly responsible for higher DPMs across manufacturers.*

Comparing Waymo to Others Using Fault Categorization. As stated earlier, we observe that AV prototypes from Waymo perform significantly better than those of its competitors. Our fault categorization allows us to speculate on reasons for this behavior. We observe (see Fig. 6) that Waymo reports significantly higher percentages of disengagements related to system faults (i.e., software or hardware issues) than machine learning/design issues, unlike other manufacturers. Extensive on-road testing (over 1,060,200 cumulative autonomous miles, which is $\sim 70\times$ more than any other manufacturer) has allowed Waymo to eliminate many fault scenarios relating to perception and control. *Even though Waymo has resolved key control and decision-making issues in the machine learning system, perception and system issues still dominate. We observe that most accidents are the result of poor decisions made by the machine learning system in complex traffic scenarios, as shown in the two case studies (in Section II). Faults in the perception systems often propagate to the decision system, leading to complex failure scenarios.* We explore this further in Section V-B, where we deal with accidents.

Table V
DISTRIBUTION OF DISENGAGEMENTS ACROSS MANUFACTURERS (AS PERCENTAGES) CATEGORIZED BY MODALITY.

Manufacturer	Automatic	Manual	Planned
Benz	47.11	52.89	0
Bosch	0	0	100
GMCruise	0	0	100
Nissan	54.2	45.8	0
Tesla	98.35	1.65	0
Volkswagen	100	0	0
Waymo	50.32	49.67	0

Last, we consider disengagements by *modality of occurrence*, i.e., whether the disengagement was initiated automatically by the AV, or manually by the driver, or as part of a planned fault injection campaign. Table V lists the distribution of these modalities across multiple manufacturers. We observe that an average of 48% of all disengagements are initiated automatically by the system. Note that this measurement is biased by manufacturers like Mercedes-Benz and Waymo that report a larger number of disengagements.

3) Question 3: Dynamics of AV Disengagements

As suggested by Fig. 5, we expect that AV technology (including perception, decision, and control) gets tuned over time, resulting in decreasing DPMs. This hypothesis is true to varying degrees across manufacturers. In this section, we further assess its validity. In particular we look at 1) the temporal dynamics of DPMs (i.e., does DPM decrease with time?), and 2) the dynamics of DPM with the cumulative number of miles driven (i.e., does DPM decrease with more extensive testing?).

Temporal Trends. Fig. 7 illustrates the temporal dynamics of the distribution of DPM per car across manufacturers aggregated per year. First, we observe that there is a distinct decreasing trend for the median DPM across most manufacturers. Some manufacturers, like Bosch that show an increase in median DPM per year claim that their disengagements result from planned fault injection experiments (see Table V). In fact, some manufacturers show a decrease of as much as $10\times$ in median DPM across the three-year analysis window. Second, we see a significant increase in the variance of the DPM across cars over the period of interest. *This increase suggests that the median performance improves over time. However, the worst-case performance does not, since the variance relative to the median is large*. In fact, for some manufacturers, like Delphi, the 75^{th} percentile DPM across years changes by less than 50%. Waymo is an exception to this trend, demonstrating a nearly $8\times$ decrease in median DPM with a significant decrease in variance across the three years of measurement. Recall from Question 1 that Waymo is still not approaching the asymptote.

Trend with Cumulative Miles Driven. While the temporal trends are important, an alternative approach is to look at disengagements per mile as a function of miles driven. Since manufacturers do not all drive the same number of autonomous miles each month, this measure is a more equitable analysis of the AVs across manufacturers. Aggregating across all manufacturers, we observe that there is a strong negative correlation between DPM and cumulative miles driven (as shown in Fig. 8). We observe that the $log(DPM)$ and

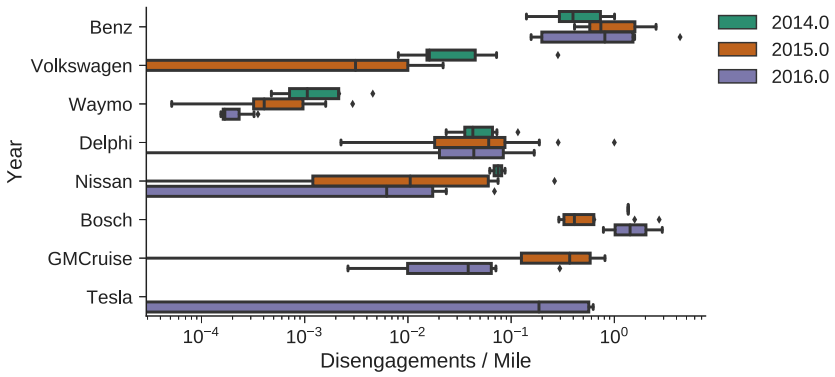


Figure 7. Time evolution (aggregated by year) of the distributions of DPMs per car across all manufacturers. The boxes show quartiles, notches show medians, and whiskers show max/mins.

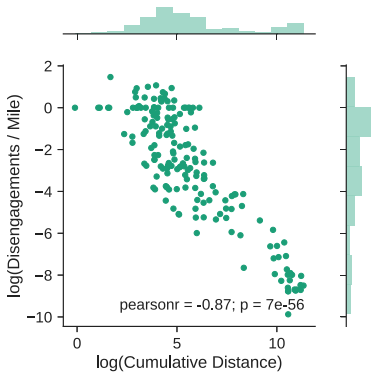


Figure 8. Linear statistical relationship between DPM per car and the cumulative number of autonomous miles.

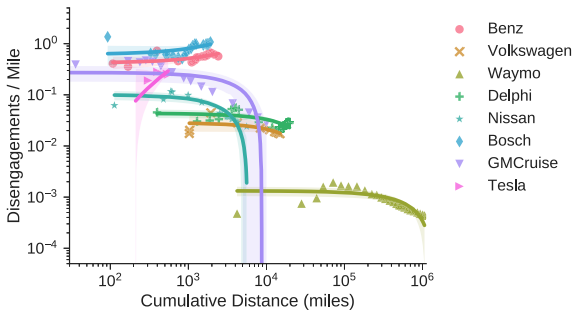


Figure 9. Evolution of DPM (per car) with the number of cumulative autonomous miles driven across all cars of that manufacturer. Lines represent a linear regression fit of each manufacturer's data.

log(cumulative autonomous miles) are correlated with a Pearson coefficient of -0.87 (at a p-value of 7×10^{-56}). Fig. 9 shows this relationship across different manufacturers, with linear regression fit lines describing the trends mentioned above. That suggests that the manufacturers are continuously improving their ADSs, with some manufacturers making more headway than others (as represented by the slope of the fitted lines). Further, we observe that manufacturers with larger DPMs seem to make more significant improvements over the same number of miles driven; this suggests that some of the faults/problems fixed as a result of this testing represent the "low-hanging fruit."

While the temporal trends maybe more indicative of how actual users will drive these cars (i.e., the AVs will be used with a mix of idle and driving times), the trends with cumulative miles provide a more robust alternative for comparisons, wherein the miles driven are the only basis for comparison. Both show a decreasing trend the first shows an increasing variance; neither shows that any of the cars have approached a very low or zero DPM regime.

4) Question 4: Driver Alertness Level

The CA DMV defines *reaction time* as "the period of time elapsed from when the autonomous vehicle test driver was alerted of the technology failure, and the driver assumed manual control of the vehicle".[6] The case studies we presented in Section II highlight the need for the human driver in the AV to be alert and cognizant of the environment. The reaction times provide an understanding of how quickly an individual would react to a fault, and hence are essential for accident avoidance. Fig. 10 gives the distribution of test drivers' reaction times across all manufacturers. We observe an average $0.85\ s$ reaction time across all test vehicle drivers and all manufacturers. This observation is consistent with a similar observation made in [34]. Further, the distribution of reaction times is long-tailed. For example, Volkswagen reported at least one case with a near $4\ hr$ reaction time for a disengagement; we suspect that this is an incorrect measurement, but cannot confirm. Fig. 11 shows this long-tailed behavior with an Exponential-Weibull fit for the reported data for manufacturers other than Volkswagen.

Comparison to Human Alertness Levels. To understand whether that behavior is indeed representative of human alertness levels when driving, we compare those results with those presented in [35] for non-AVs. [35] found the reaction time for braking in test vehicles to be $0.82\ s$. This observation is consistent with our study. Further, [35] report that a driver's ownership of a vehicle (i.e., it is his or her own property) increased reaction time by approximately $0.27\ s$.

[6]We assume the reaction times to be upper bounded where they are listed as ranges.

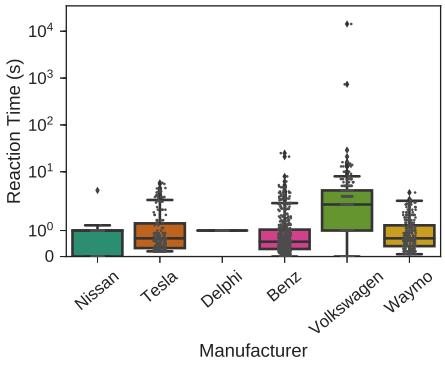
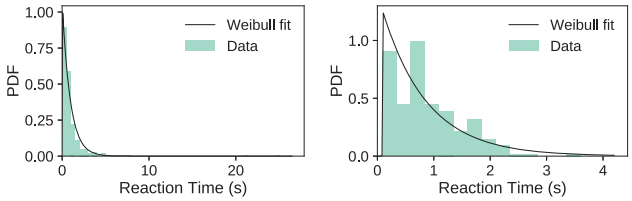


Figure 10. Distribution of reaction times for drivers in case of a disengagement across all manufacturers. The boxes show quartiles, notches show medians, and whiskers show max/mins.



(a) Mercedes-Benz (b) Waymo

Figure 11. Distribution of reaction times for the Mercedes-Benz and Waymo.

Hence we assume $1.09\ s$ to be the average time for a human driver in a non-AV to respond any situation on the road. The observation implies that semi-AVs which are the most commonly deployed AVs on public streets) would require continued human supervision and alertness similar to human controlled non-AVs. Echoing the results of Question 3, that in turn suggests that the technology may not be mature enough to allow human drivers to be engaged in other activities, contrary to what is advertised.

Temporal Behavior of Reaction Time. We find that a driver's alertness decreases (i.e., reaction time increases) with the number of cumulative miles driven. At a 99% confidence level, we observe a positive correlation between the cumulative miles driven and the reaction times across manufacturers. For example, Waymo and Mercedes-Benz show a Pearson's correlation coefficient of 0.19 (at p-value $= 0.01$) and 0.11 (at p-value $= 0.007$), respectively. Taken together, that observation and the previous observation about decreasing DPM (described in Section V-A3) suggest that a driver's alertness decreases as the system's performance improves (i.e., DPM decreases).

Fault Detection Latency and Reaction Time. By definition, the reaction time does not include fault detection time. However, as our case studies show, the detection time is indeed part of the end-to-end time window in which the driver reacts to an adverse situation. For example, in both case studies presented in Section II, the primary cause of the accident was the insufficient time left for the driver to make a decision after the fault was detected.

The drivers of AVs have to maintain the same level of alertness as when driving non-AVs. This suggests that the small size of the overall action window (detection time + reaction time) can make the reaction-time-based accidents a frequent failure mode with the widespread deployment of AVs. We also note that in planned test scenarios for AVs, drivers are required, trained, and paid to remain continuously attentive to the activities of the AV. Data for them might not generalize to regular users.

B. Analysis of AV Accident Reports

1) Question 5: Comparison to Human Drivers

To address this question, we define two additional measures: 1) accidents per mile (APM), and 2) disengagements per accident (DPA). We calculate the DPAs as shown in Table VI. As some of the accident reports were partially redacted by the CA DMV to obfuscate AV identification (e.g., the registration number or VIN number were removed), we cannot compute the APM per vehicle directly. We instead compute *accidents per mile* using the equation $APM = {}^{DPM}/_{DPA}$. Even though the number of accidents is small compared to the number of disengagements, we use [36] to test the statistical significance of our results. Our calculations for two out of the 4 manufacturers (i.e., Waymo and GMCruise) were made at $> 90\%$ significance.

Table VI
SUMMARY OF ACCIDENTS REPORTED BY MANUFACTURERS.

Manufacturer	Accidents	Fraction of Total	DPA
Waymo	25	59.52	18
Delphi	1	2.38	572
Nissan	1	2.38	135
GMCruise	14	33.33	20
Uber ATC	1	2.38	–

DPA = Disengagements per accident.

Comparison of APMs across Manufacturers. We observe that there is great variability ($\sim 100\times$) in APMs across manufacturers (see Table VII). For example, Waymo is responsible for 59.52% of accidents reported (see Table VI), but has the lowest DPM (7.45×10^{-4}), the lowest DPA (18), and the lowest APM (4.14×10^{-5}). In contrast, GMCruise has a similar DPA (20) but performs $238\times$ worse in terms of DPM, and $214\times$ worse in terms of APM, as compared to Waymo (see Table VII). This suggests that there is significant variability across manufacturers in classifying the severity of disengagements, which again indicates the immaturity of the current AV technology. Also, the observed APM metric variability can be partially attributed to test drivers' proactive disengagement of the ADS (i.e., manual disengagement as presented in Section V-A2) to prevent accidents. We compare the accident rate of AVs with that of manual vehicles using data for [37], [38], which report that one accident is expected every $500,000$ miles (i.e., $APM = 2 \times 10^{-6}$). We find that compared to human drivers, AVs perform 15–$22\times$ worse (see Table VII) in terms of APM.[7]

When they are calculated using first principles (i.e., not using DPA as done before), for vehicles that can be identified in the accident reports, we observe a strong positive correlation between the number of accidents observed per mile and the number of autonomous miles driven (with a Pearson correlation coefficient of 0.98 at p-value < 0.01). Comparing that number to the trends in the DPM seen in Fig. 8, we see that there is a much stronger correlation of the APM with cumulative miles. This behavior might be indicative of the manufacturers' priority on fixing problems in their ADSs (i.e., they identify problems relating to accidents and fix them quickly).

Our analysis shows that for the same number of miles driven, for manufacturers that reported accidents, human-driven cars (non-AVs) are $15 - 4000\times$ less likely to have an accident than AVs.

Collision Speeds and Locations. All the accidents reported in the dataset occurred at low speeds and in the vicinity of intersections on urban streets. Fig. 12 shows that more than 80% of the accidents occurred when the relative speed[8] of the colliding vehicles was less than $10\ mph$. In most of the cases in which the non-AV vehicle was determined to be at fault, the underlying cause can be attributed to the failure of the vehicle's driver to anticipate AV behavior. This observation points to

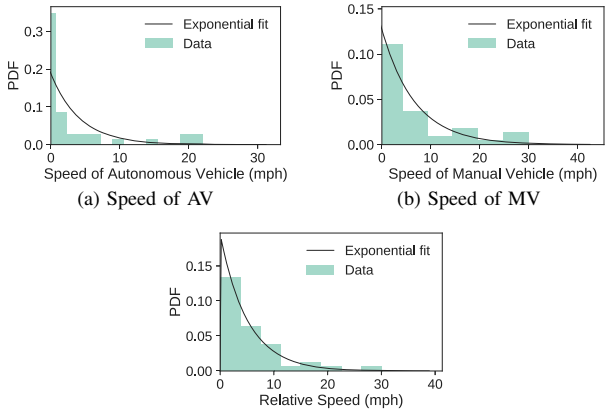
[7]Note that [37], [38] report only crashes on highways and freeways. However, AVs are required to report any crash on all types of roads.

[8]The absolute difference between the speeds of the vehicles at the collision.

Table VII
RELIABILITY OF AVS COMPARED TO HUMAN DRIVERS.

Manufacturer	Median DPM (mile^{-1})	Median APM (mile^{-1})	Rel. to HAPM
Mercedes-Benz	0.565	–	–
Volkswagen	0.0181	–	–
Waymo	0.000745	4.140×10^{-5}	20.7×
Delphi	0.0263	4.599×10^{-5}	22.99×
Nissan	0.0413	3.057×10^{-4}	15.285×
Bosch	0.811	–	–
GMCruise	0.177	8.843×10^{-3}	4421.5×
Tesla	0.250	–	–

HAPM – Human APM.
Human APM $= 2 \times 10^{-6}$mile^{-1} [37], [38].
Column 4 = $^{\text{AV APM}}/_{\text{Human APM}}$.



(a) Speed of AV
(b) Speed of MV
(c) Relative speed

Figure 12. Distribution of vehicular speeds for all reported accidents.

the need for better understanding of the driving interactions and behaviors that drivers expect from other on-road vehicles. Most of the accidents were minor (either rear-end or side-swipe collisions), and no serious injuries were reported.

Our data show that better situational awareness needs to be provided by the ADSs (in particular the machine learning algorithms) to preemptively avoid accidents in a timely fashion.

C. Discussion

1) Comparison to Other Safety-critical Autonomous Systems

Airplanes [39] and surgical robots [40] are safety-critical semi-autonomous systems that have seen ubiquitous deployment, as well as a significant body of work characterizing and improving their resilience. We compare AVs to both of these systems in terms of the accidents per mission (APMi), to gauge the maturity of AVs vis-a-vis these systems. We define a *mission* as the continuous operation of the system of interest from the time of commencement to the end of the activity. For airplanes and cars, a mission is equivalent to one departure (i.e., trip), and for the surgical robot, a mission is equivalent to a surgical procedure.

We use data presented in [41] (9.8 accidents per $100,000$ departures for airplanes) and [42] (1043 accidents per $100,000$ procedures for surgical robots) as the baseline for comparison. We estimate the APMi of an AV by using data (pertaining to the average length of a vehicle ride on U.S. public roads for which there is a median of 10 miles per trip) presented in [43]. Using the APM metric computed earlier as shown in Table VIII, we compute APMi as APM $\times$ length of the

Table VIII
RELIABILITY OF AVS COMPARED TO OTHER SAFETY-CRITICAL AUTONOMOUS SYSTEMS.

Manufacturer	APMi	Aviation Industry $^{\text{APMi}}/_{\text{Airline APM}}$ [41]	Surgical Robotics $^{\text{APMi}}/_{\text{SR APM}}$ [42]
Waymo	4.140×10^{-4}	4.22	0.0398
Delphi	4.599×10^{-4}	4.69	0.0442
Nissan	3.057×10^{-3}	31.19	0.293
GMCruise	8.843×10^{-2}	902.34	8.502

APMi = Accidents per mission for an AV
Airline APM = 9.8×10^{-5}
Surgical Robot (SR) APM = 1.04×10^{-2}

average trip. Our analysis shows that AVs do surprisingly well per mission. *Compared to airplanes (which utilize sophisticated resilience models and techniques), AVs are merely* $4.22\times$ *worse, and are* $2.5\times$ *better than surgical robots (see Table VIII).*

However, if all cars are replaced by AVs in the future, the AVs will make ~ 96 billion trips per year [44], compared to the 9.6 million trips for airlines. This means that AVs will make $10,000\times$ more trips than airlines, leading to a higher number of accidents per year than for airplanes. Further, the average length of a mission in terms of time and miles covered is significantly different for airplanes and AVs. Hence a holistic comparison across these systems would need to consider operational time per mission, as well as account for competing failures across concurrent deployments of these systems.

2) Traditional Reliability Metrics

While we have made an approximate comparison above, the more traditional and accurate method for comparing the resilience of AVs with that of airplanes (which are also highly automated systems) is via operational hours to failure. That metric, however, is unavailable for cars, since we do not have information about the idle time for these vehicles or its distribution. We propose an alternative metric based on the number of miles driven to disengagement/accident. This metric will be available across transportation systems.

To directly obtain this measure, there needs to be a small change in the data collection by the DMV: manufacturers and the DMV should collect data on miles between disengagements per vehicle to enable the computation of the metrics.

VI. THREATS TO VALIDITY

An empirical study like ours is subject to vagaries arising from heterogeneous data collection systems (e.g., the inclusion or exclusion of data points, or the disparate information content across data formats), thus hampering the ability to draw generalized conclusions. Dealing with such issues is not uncommon in the realm of system reliability assessment. We assert the need for replication studies to verify our conclusions across other datasets. We now discuss potential threats to validity that are specifically related to our study.

Construct Validity implies that variables associated with the study are measured correctly, i.e., that the measurements are constructed in accordance with the theoretical foundations of the area. We have discussed construct validity in Section V-C2.

Internal Validity implies that there are no systematic errors and biases. We studied the datasets available from 12 different manufacturers and only reported generalized trends in order

to eliminate any biases and micro-observations (observations with low statistical significance) that might be artifacts of bad logging or biases from the manufactures in reporting the disengagements and accidents. For example:

- *Data underreporting*: In order to obtain an AV testing permit, companies are legally required to catalogue and submit to the DMV reports of all disengagements and accidents that 1) pertained to technology failures and safe operation of the AVs, and 2) required the AV test driver to disengage the autonomous mode and take immediate manual control of the vehicle. The interpretation of "safe" operation and technology "failure" can vary across manufacturers, leading to underreporting. Further, regulatory oversight and enforcement of regulations are difficult and may result in underreporting. Given the available data, we cannot accurately estimate the scale of underreporting, and hence refrain from drawing any such conclusions.
- *Not all miles are equivalent*: One manufacturer may hold the tests of its AVs in more challenging environments than others do, e.g., at night or during bad weather. Not all manufacturers report environmental conditions during tests. Where available, we report the testing conditions and disengagements caused by environmental factors (see "Environment" in Fig. 6).
- *Validity of fault tags and failure categories*: There is no consistent data format for the provided disengagement/accident reports across manufacturers. Our NLP framework for tagging and categorization may lead to systematic errors; therefore, the dictionaries were verified manually by the authors to ensure their correctness. We explicitly labeled data points as "Unknown-T/C" when there was uncertainty in the tags and categories given by the NLP framework.

External validity concerns the extent to which a study can be generalized to other systems or datasets. To the best of our knowledge, the CA DMV dataset is the only publicly available dataset pertaining to AV failures. Until we work with manufacturers on proprietary data (which might not be disclosed publicly), we cannot comment on the general external validity of the techniques presented here.

VII. Related Work

The majority of the prior research into AV systems focuses on the functionality of vehicle guidance systems. Numerous demonstrations of end-to-end computing systems for autonomous vehicles have recently been done (e.g., [2]–[7], [45], [46]). The currently accepted practice for vehicular safety, based on the ISO 26262 safety standard [47], is to consider human drivers to have ultimate responsibility for safety. That is the basis for most AV testing programs on public roads, which require a safety driver to be in the vehicle to monitor the vehicle. This driver is expected to intervene if a system failure occurs that leads to a disengagement or accident; indeed, we observe several such incidents in the CA DMV datatsets. In such a scenario, safety considerations for the AV are driven by 1) the AV's ability to alert the driver in case of failure, 2) the driver's ability to recognize the abilities of the AV and the limits of the system, 3) the AV's ability to anticipate the behavior of other road users who might not always conform to the rules, and 4) the other road user's ability to anticipate the behavior of the AV [48], [49]. How this will be handled in autonomous vehicles remains an open question [50]. Safety is also emphasized in a number of publications, including [51], [52]. Waymo has published a report on the safety precautions considered for their AVs [25].[9]

[36] provides a model to estimate the number of miles AVs have to be driven to demonstrate their reliability with statistical confidence. [30], [34] provide summary statistics (e.g., driver reaction times and AV speed in accident scenarios) from tabulated data in the DMV dataset. Our approach uses an STPA based ontology and NLP techniques (which in itself are novel contributions of this work) to parse a significant amount of unstructured data presented as natural text.

[19] use fault injection to evaluate the fault tolerance of deep neural networks (DNN: used primarily in the *Sensor Fusion & Environmental Information Processing* step shown in Fig. 3), analyze the DNN's results, and propose techniques to safeguard DNNs from single-event upsets. In contrast, we present an analysis of the entire control system of the AV, of which DNNs are a small part.

Other related work has focused on safety and reliability of AVs as they apply to legal (e.g., [10], [11]) and regulatory barriers (e.g., [12]–[15]) for AV deployment and implementation.

Security and privacy measures to encompass system-level attacks and failures of AVs have also been studied [53], [54].

VIII. Conclusions and Future Work

A steady march toward the use of AVs is clearly under way. The reliability and safety challenges of fully-autonomous vehicles (Level 4 & 5, currently under development) and today's semi-AVs are significant and underestimated. We therefore draw the following conclusions to frame our future research and draw the attention of other reliability researchers.

- There is ongoing research on the verification and validation of the safety properties of individual system components (e.g., the control, communication, and mechanical system components) using the STAMP framework [51]. However, our study shows there is a need for rigorous theoretical models (like STPA models) for evaluating AV technologies.
- The machine learning systems responsible for perception and control need further research and assessment under fault conditions via stochastic modeling and fault injection to augment data collection.
- In reality, there is a strong possibility that both AVs and semi-AVs will co-exist with non-AVs (with human drivers completely in charge) within several years. Therefore the urgency of joint study driven by data and models needs to be emphasized.

Acknowledgments

This material is based upon work supported by an IBM Faculty Award, and by the National Science Foundation (NSF) under Grant Nos. CNS 13-14891 and CNS 15-45069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. We thank K. Atchley, J. Applequist, W. S. B. M. Lim, and A. P. Athreya for their help in preparing the manuscript.

[9]For their trained drivers, Waymo claimed there was 1 accident for 2.3 million miles; we cannot substantiate that.

REFERENCES

[1] M. Gerla *et al.*, "Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds," in *2014 IEEE World Forum on Internet of Things (WF-IoT)*, Mar 2014, pp. 241–246.

[2] U. Ozguner, C. Stiller, and K. Redmill, "Systems for safety and autonomous behavior in cars: The DARPA grand challenge experience," *Proc. of the IEEE*, vol. 95, no. 2, pp. 397–412, Feb 2007.

[3] "Special issue on the 2007 DARPA Urban Challenge, Part I," *J. Field Robot.*, vol. 25, no. 8, Aug 2008.

[4] C. Urmson *et al.*, "Autonomous driving in urban environments: Boss and the urban challenge," *J. Field Robotics*, vol. 25, no. 8, pp. 425–466, 2008.

[5] A. Chatham, "Google's self-driving cars—the technology, capabilities, and challenges," in *2013 Embedded Linux Conf., Feb*, 2013, pp. 20–24.

[6] C. Urmson, "Realizing self-driving vehicles," in *2012 IEEE Intelligent Vehicles Symposium (IV). Alcalá des Henares, Spanien*, 2012.

[7] B. Paden *et al.*, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Trans. Intelligent Vehicles*, vol. 1, no. 1, pp. 33–55, Mar 2016.

[8] W. Payre, J. Cestac, and P. Delhomme, "Intention to use a fully automated car: Attitudes and a priori acceptability," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 27, pp. 252–263, Nov 2014.

[9] S. Shladover, D. Su, and X.-Y. Lu, "Impacts of cooperative adaptive cruise control on freeway traffic flow," *Transportation Research Record: J. the Transportation Research Board*, vol. 2324, pp. 63–70, Dec 2012.

[10] G. E. Marchant and R. A. Lindor, "The coming collision between autonomous vehicles and the liability system," *Santa Clara L. Rev.*, vol. 52, p. 1321, 2012.

[11] M. Parent *et al.*, "Legal issues and certification of the fully automated vehicles: best practices and lessons learned," *CityMobil2 Rep.*, 2013.

[12] J. M. Anderson *et al.*, "Autonomous vehicle technology: A guide for policymakers," RAND Corp., Tech. Rep. RR-443-2-RC, 2016.

[13] D. J. Fagnant and K. Kockelman, "Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations," *Transportation Research Part A: Policy and Practice*, vol. 77, pp. 167–181, Jul 2015.

[14] L. Fraade-Blanar and N. Kalra, "Autonomous vehicles and federal safety standards: An exemption to the rule?" RAND Corp., Tech. Rep. PE-258-RC, 2017.

[15] D. G. Groves and N. Kalra, "Enemy of good," RAND Corp., Tech. Rep. RR-2150-RC, 2017.

[16] California Department of Motor Vehicles, "Testing of autonomous vehicles," https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/testing, Accessed: 2017-11-27.

[17] C. Chen *et al.*, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE International Conf. Computer Vision*, 2015, pp. 2722–2730.

[18] S. Petti and T. Fraichard, "Safe motion planning in dynamic environments," in *2005 IEEE/RSJ International Conf. Intelligent Robots and Systems*, Aug 2005, pp. 2210–2215.

[19] G. Li *et al.*, "Understanding Error Propagation in Deep Learning Neural Network (DNN) Accelerators and Applications," in *Proc. International Conf. for High Performance Computing, Networking, Storage and Analysis*, 2017, pp. 8:1–8:12.

[20] NVIDIA, "Introducing Xavier, the NVIDIA AI Supercomputer for the Future of Autonomous Transportation," https://blogs.nvidia.com/blog/2016/09/28/xavier, Accessed: 2017-11-27.

[21] SAE International, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, Sep 2016.

[22] F. Mujica, "Scalable electronics driving autonomous vehicle technologies," *Texas Instruments White Paper*, 2014.

[23] N. Leveson, *Engineering a safer world: Systems thinking applied to safety*. MIT press, 2011.

[24] A. Abdulkhaleq *et al.*, "A Systematic Approach Based on STPA for Developing a Dependable Architecture for Fully Automated Driving Vehicles," *Procedia Engineering*, vol. 179, pp. 41–51, 2017.

[25] Waymo Safety Report, "On the Road to Fully Self-Driving," https://assets.documentcloud.org/documents/4107762/Waymo-Safety-Report-2017.pdf, Accessed: 2017-11-27.

[26] N. H. Amer *et al.*, "Modelling and control strategies in path tracking control for autonomous ground vehicles: a review of state of the art and challenges," *J. Intelligent & Robotic Systems*, vol. 86, no. 2, pp. 225–254, 2017.

[27] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The Kitti Vision Benchmark Suite," in , *2012 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361.

[28] California Department of Motor Vehicles, "Deployment of autonomous vehicles for public operation," https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/auto, Accessed: 2017-11-27.

[29] F. M. Favarò *et al.*, "Examining accident reports involving autonomous vehicles in california," *PLoS one*, vol. 12, no. 9, p. e0184952, 2017.

[30] F. Favarò, S. Eurich, and N. Nader, "Autonomous vehicles' disengagements: Trends, triggers, and regulatory limitations," *Accident Analysis & Prevention*, vol. 110, pp. 136–148, 2018.

[31] R. Smith, "An Overview of the Tesseract OCR Engine," in *Ninth International Conf. Document Analysis and Recognition*, vol. 2, Sep 2007, pp. 629–633.

[32] H. Alemzadeh, "Data-driven resiliency assessment of medical cyber-physical systems," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2016.

[33] A. Avizienis *et al.*, "Basic concepts and taxonomy of dependable and secure computing," *IEEE Trans. Dependable Secur. Comput.*, vol. 1, no. 1, pp. 11–33, Jan. 2004.

[34] V. V. Dixit, S. Chand, and D. J. Nair, "Autonomous vehicles: disengagements, accidents and reaction times," *PLoS one*, vol. 11, no. 12, p. e0168054, 2016.

[35] D. B. Fambro, *Determination of stopping sight distances (Report / National Cooperative Highway Research Program)*. National Academy Press, 1997.

[36] 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.

[37] National Highway Traffic Safety Administration (NHTSA), "2015 motor vehicle crashes: overview DOT HS 812 318," *Traffic Safety Facts Research Note*, pp. 1–9, 2016.

[38] Federal Highway Administration (FHWA),, "Traffic volume trends." https://www.fhwa.dot.gov/policyinformation/travel_monitoring/tvt.cfm, Accessed: 2017-11-27.

[39] FAA, "System design and analysis," Tech. Rep. AC 25.1309-1A, Jun 1988.

[40] H. Alemzadeh *et al.*, "Adverse events in robotic surgery: a retrospective study of 14 years of FDA data," *PLoS One*, vol. 11, no. 4, p. e0151470, 2016.

[41] National Transportation Safety Board, "Aviation statistics: Review of accident data," https://www.ntsb.gov/investigations/data/Pages/aviation_stats.aspx, Accessed: 2017-11-27.

[42] H. Alemzadeh *et al.*, "Analysis of safety-critical computer failures in medical devices," *IEEE Security & Privacy*, vol. 11, no. 4, pp. 14–26, Jul 2013.

[43] U.S. Department of Transportation, Federal Highway Administration, Office of Highway Policy Information, National Household Travel Survey., "Our nation's highways: 2008," https://www.fhwa.dot.gov/policyinformation/pubs/pl08021/index.cfm, Accessed: 2017-11-27.

[44] P. Plötz, N. Jakobsson, and F. Sprei, "On the distribution of individual daily driving distances," *Transportation Research Part B: Methodological*, vol. 101, pp. 213–227, 2017.

[45] G. Stanek *et al.*, "Junior 3: A test platform for advanced driver assistance systems," in *Intelligent Vehicles Symposium (IV), 2010 IEEE*, 2010, pp. 143–149.

[46] J. Levinson *et al.*, "Towards fully autonomous driving: Systems and algorithms," in *2011 IEEE Intelligent Vehicles Symposium (IV)*, Jun 2011, pp. 163–168.

[47] "Road vehicles — Functional safety ," International Organization for Standardization, Geneva, CH, Standard, Nov. 2011.

[48] S. M. Casner, E. L. Hutchins, and D. Norman, "The challenges of partially automated driving," *Comm. of the ACM*, vol. 59, no. 5, pp. 70–77, Apr 2016.

[49] A. Reschka, "Safety concept for autonomous vehicles," in *Autonomous Driving: Technical, Legal and Social Aspects*, M. Maurer *et al.*, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 473–496.

[50] P. Koopman and M. Wagner, "Autonomous vehicle safety: An interdisciplinary challenge," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 1, pp. 90–96, Spring 2017.

[51] N. Leveson, "A new accident model for engineering safer systems," *Safety science*, vol. 42, no. 4, pp. 237–270, 2004.

[52] C. Fan *et al.*, "DryVR: Data-Driven Verification and Compositional Reasoning for Automotive systems," in *Computer Aided Verification*. Springer International Publishing, 2017, pp. 441–461.

[53] SAE International, *Cybersecurity Guidebook for Cyber-Physical Vehicle Systems*, Jan 2016.

[54] J. Joy and M. Gerla, "Privacy risks in vehicle grids and autonomous cars," in *Proc. of the 2nd ACM International Workshop on Smart, Autonomous, and Connected Vehicular Systems and Services*, 2017, pp. 19–23.