My Precious Crash Data: Barriers and Opportunities in Encouraging Autonomous Driving Companies to Share Safety-Critical Data

HAUKE SANDHAUS, Cornell University, Cornell Tech, USA
ANGEL HSING-CHI HWANG, Cornell University, USA
WENDY JU, Cornell Tech, USA
QIAN YANG, Cornell University, USA

Safety-critical data, such as crash and near-crash records, are crucial to improving autonomous vehicle (AV) design and development. Sharing such data across AV companies, academic researchers, regulators, and the public can help make all AVs safer. However, AV companies rarely share safety-critical data externally. This paper aims to pinpoint why AV companies are reluctant to share safety-critical data, with an eye on how these barriers can inform new approaches to promote sharing. We interviewed twelve AV company employees who actively work with such data in their day-to-day work. Findings suggest two key, previously unknown barriers to data sharing: (1) Datasets inherently embed salient knowledge that is key to improving AV safety and are resource intensive. Therefore data sharing, even within a company, is political and fraught. (2) Interviewees believed AV safety knowledge is private knowledge that brings competitive edges to their companies, rather than public knowledge for social good. vWe discuss the implications of these findings for incentivizing and enabling safety-critical AV data sharing, specifically, implications for new approaches to (1) debating and stratifying public and private AV safety knowledge, (2) innovating data tools and data sharing pipelines that enable easier sharing of public AV safety data *and knowledge*; (3) offsetting costs of curating safety-critical data and incentivizing data sharing.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Autonomous Driving, Data Work, Safety, Industry Practice.

1 INTRODUCTION

Sharing data of crashes and near-crashes holds great potential to improve autonomous vehicle (AV) safety research and oversight [45, 55, 61, 63]. Without sufficient safety-critical data, AV safety performance could drop significantly. For example, prior research shows insufficient corner cases in the training data could cause the accuracy of object detection models to drop to 12.8% mean Average Recall, resulting in unsafe driving conditions [48]. This led to the release of CODA, a public dataset of real-world road corner cases for object detection in autonomous driving. Similar movements toward open-sourcing safety-critical data allow researchers across the industry and academia to jointly investigate the causes of hazardous AV driving conditions. The increased availability of such data can also facilitate AV designers and developers to conceive preventative solutions before life-threatening accidents occur.

To date, AV companies rarely share safety-critical data externally. While policies mandate sharing specific types of AV safety-critical data, companies rarely share beyond the minimal requirements [19, 30]. Recognizing this problem, grassroots movements have started crowdsourcing information about AV crashes and near-crashes [3]. AV safety researchers have started curating and sharing data of simulated crashes [22, 34, 41]. Others have started developing data tools that make data sharing easier [9, 21, 79], These approaches have been highly valuable and impactful. Yet questions remain:

Authors' addresses: Hauke Sandhaus, hgs52@cornell.edu, Cornell University, Cornell Tech, 2 West Loop Rd, New York, New York, USA, 10044; Angel Hsing-Chi Hwang, hh695@cornell.edu, Cornell University, Ithaca, New York, USA, 14853; Wendy Ju, wendyju@cornell.edu, Cornell Tech, 2 West Loop Rd, New York, New York, USA, 10044; Qian Yang, qianyang@cornell.edu, Cornell University, Ithaca, New York, USA, 14853.

Why haven't AV companies started to share their safety-critical data externally and systematically, given the now available data-sharing tools? What other approaches might get companies to do so?

This paper aims to pinpoint why AV companies are reluctant to share safety-critical data, with an eye on how these barriers can inform new approaches to promote sharing. Toward this goal, we interviewed twelve AV company employees who work with safety-critical data for AV design and deployment in their day-to-day work. The interviews focused on understanding their current data management and sharing practices, the challenges and concerns for safety-critical data sharing they have encountered, and the ideal data sharing practices they wish for.

Our interviews identified two key, previously unknown barriers to AV data sharing. Both underscore that AV companies' lack of incentive to share data—more so than the pragmatic challenges around how to share it—as a primary reason behind the rarity of data sharing. First, an AV company's crash and near-crash data inherently embed knowledge about the machine learning (ML) models and infrastructure that the company uses to *improve* AV safety. Therefore, such datasets are resource- and knowledge-intensive to curate. Data sharing, even within a company, is political and fraught. Second, interviewees believed AV safety knowledge is private knowledge that brings competitive edges to their companies, rather than public knowledge for social good.

Re-framing the challenges of AV safety data-sharing as a problem of incentives (why share?) rather than a problem of tools (how to share?) illuminates new approaches to addressing these challenges. We see a need for AV safety-related communities—academics, policymakers, AV companies, the general public, etc.—to debate and stratify public and private AV safety knowledge. For example, what AV crash data must be shared with regulatory agencies or other AV companies so that similar accidents will not occur? We see an opportunity for researchers and practitioners to innovate data tools and data-sharing pipelines that make it easier to stratify public and private knowledge in AV datasets. For example, we see an opportunity to build shared Virtual Reality environments that encode crash-prone road scenarios, allowing companies to share knowledge/data about safety-critical *situations* without sharing their preparatory ML models that handle these situations. Finally, we see an opportunity for academics and policymakers to innovate strategies for incentivizing data sharing, to make AV safety data curation and share at least partially a cause for public good, rather than an issue of sheer capitalistic competition. This paper discusses these potential new approaches to improving data sharing, drawn from our interview findings.

We make two contributions through the present research. To begin with, we delineate the fundamental causes that hinder sharing of safety-critical AV data. This informs more targeted approaches to motivating data-sharing practice. We highlight three plausible means: defining public vs. private data knowledge, redirecting the design goals of data-sharing tools, and executing incentive programs. Grounded on these actionable proposals, we call for attention and input from the AutoUI and Human-Computer Interaction communities to make cross-industry-academia data-sharing practices more commonplace.

2 RELATED WORK

At the outset of this paper, we plunged head first into the notion of "*AV safety-critical data*" without defining the term. There is no one agreed-upon, precise definition of AV safety-critical data across existing literature. Instead, researchers have used the term to broadly refer to data recorded from AV crashes and near-crash events. These incidents can occur when various instances of unseen objects, circumstances, and behaviors occur, such as drivers violating traffic rules, demonstrating unsafe driving behaviors, driving under extreme weather conditions, novel objects on roads, or less-common, out-of-context behavior by traffic participants. In this paper, we use "*AV safety-critical data*" to refer to such data as well.

2.1 Benefits of Sharing Safety-Critical Data

Availability of safety-critical data is key to facilitating AV safety research, collaboration, and oversight [1, 46, 48, 54]. The AutoUI community has advocated for AV companies to share and provide access to such data and named specific data types that can be most relevant and useful for improving AV design [24].

Existing research has shown the promise of such data for AV safety design, e.g., by adversarial generation of safety-critical events in simulation, and showing the improved performance by re-training on them [38, 75]. This data can allow an end-to-end approach to improving autonomous driving, building machine learning (ML) models that use driving context data and generate safer AV driving behaviors [16]. Moreover, AV researchers and regulators need AV safety-critical data to investigate reasons for crashes (e.g., Uber's 2018 fatal crash by Macrae [52]), assign responsibilities, and devise strategies for preventing similar incidents.

The automotive user interface and user research communities depend on real field data to further user-centered design for AVs [25], and maintaining scientific integrity necessitates adherence to open data practices [24]. This paper investigates whether and why sharing these data types remains limited despite these numerous benefits.

2.2 The Lack of Safety-Critical Data Sharing

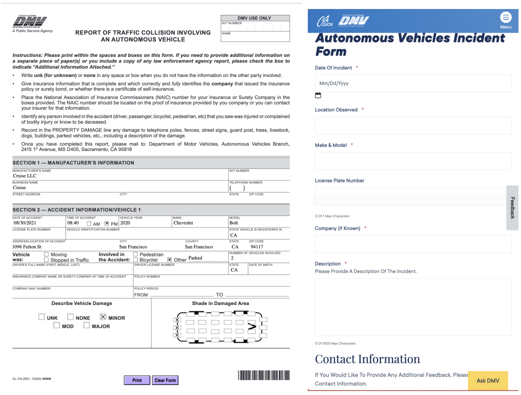
Despite the above-mentioned promises, abundant evidence shows that safety-critical data remains mostly unavailable to the greater AV research and design community [35, 48]. Several policies in place uphold this status quo. The current AV testing policies in Europe and the United States demand minimal crash test data sharing. But, such data seldom fully capture factors that cause safety-threatening events [17, 57]. Often, required data types are limited to descriptive statistics and general information, such as the month when an AV crash occurred, the manufacturer involved, and whether there were injuries.

Furthermore, officially designated means for data collection do not support recording rich forms of data. Thus, datasets published through government authorities often lack adequate details to inform AV safety design. For instance, AV companies are required to report safety-critical events through text documents (e.g., DMV OL 316 [13, 53] and the DMV autonomous vehicle incident web form [14]) in the United States. At most, these forms provide information such as crashes per mile without further details about each incident [36]. New European policies have mandated that all European-sold vehicles with higher-level automated systems include Event Data Recorders, colloquially known as 'Black boxes.' However, data collected through these devices do not capture detailed information about the location, trajectory, time, or context of safety-critical events [12, 28, 29].

Under these regulations, companies share little beyond the minimal requirements. Recent reviews suggested that datasets shared across the AV industry mostly consist of everyday driving records *only*, rather than those highlighting safety concerns [45, 47, 79]. As a result, even the most comprehensive datasets on AV crashes (see Table 3 from [81]) lack critical details about safety hazards. For example, they do not provide the exact time of the crash, information about the safety driver, speed at the moment of the crash, or micro-location and detailed time-series and movement data.

Insufficient data-sharing is reflected in two common phenomena: First, there are grassroots movements to crowdsource safety-related data. For example, websites are collecting data on deaths involving Tesla's Autopilot [3], and another tracking automation incidents [18]. Secondly, due to the lack of real safety-critical data, existing studies have resorted to using simulated data to enhance AV safety design. Past research has created crash scenarios by introducing virtual anomalies on roads [10], [51], such as the StreetHazard dataset [40], while some researchers have relied on police reports to roughly reconstructing crash scenarios [31, 32].

Together, several studies and reviews of datasets have concluded that the availability of safety-critical data is inadequate for ensuring reliable AV design [35, 45, 47, 48, 79]. An overview describing the categories of public and open-sourced safety-critical data sources available to AV researchers is provided in Table 1.



A Public Service Agency

REPORT OF TRAFFIC COLLISION INVOLVING AN AUTONOMOUS VEHICLE

DMV USE ONLY
AVT NUMBER
NAME

Instructions: Please print within the spaces and boxes on this form. If you need to provide additional information on a separate piece of paper(s) or you include a copy of any law enforcement agency report, please check the box to indicate "Additional Information Attached."

- Write **unk (for unknown)** or **none** in any space or box when you do not have the information on the other party involved.
- Give insurance information that is complete and which correctly and *fully* identifies the **company** that issued the insurance policy or surety bond, or whether there is a certificate of self-insurance.
- Place the National Association of Insurance Commissioners (NAIC) number for your Insurance or Surety Company in the boxes provided. The NAIC number should be located on the proof of insurance provided by you company or you can contact your insurer for that information.
- Identify any person involved in the accident (driver, passenger, bicyclist, pedestrian, etc) that you saw was injured or complained of bodily injury or know to be deceased.
- Record in the PROPERTY DAMAGE line any damage to telephone poles, fences, street signs, guard post, trees, livestock, dogs, buildings, parked vehicles, etc., including a description of the damage.
- Once you have completed this report, please mail to: Department of Motor Vehicles, Autonomous Vehicles Branch, 2415 1st Avenue, MS D405, Sacramento, CA 95818

SECTION 1 — MANUFACTURER'S INFORMATION

MANUFACTURER'S NAME Cruise LLC
AVT NUMBER
BUSINESS NAME Cruise
TELEPHONE NUMBER ( )
STREET ADDRESS
CITY
STATE
ZIP CODE

SECTION 2 — ACCIDENT INFORMATION/VEHICLE 1

DATE OF ACCIDENT 08/30/2021
TIME OF ACCIDENT 08:40 □ AM ☒ PM
VEHICLE YEAR 2020
MAKE Chevrolet
MODEL Bolt
LICENSE PLATE NUMBER
VEHICLE IDENTIFICATION NUMBER
STATE VEHICLE IS REGISTERED IN CA
ADDRESS/LOCATION OF ACCIDENT 1096 Fulton St
CITY San Francisco
COUNTY San Francisco
STATE CA
ZIP CODE 94117
Vehicle was: □ Moving □ Stopped in Traffic
Involved in the Accident: □ Pedestrian □ Bicyclist ☒ Other Parked
NUMBER OF VEHICLES INVOLVED 2
DRIVER'S FULL NAME (*FIRST, MIDDLE, LAST*)
DRIVER LICENSE NUMBER
STATE CA
DATE OF BIRTH
INSURANCE COMPANY NAME OR SURETY COMPANY AT TIME OF ACCIDENT
POLICY NUMBER
COMPANY NAIC NUMBER
POLICY PERIOD FROM ______ TO ______

Describe Vehicle Damage
□ UNK □ NONE ☒ MINOR
□ MOD □ MAJOR

Shade in Damaged Area

OL 316 (REV. 7/2020) **WWW**

Autonomous Vehicles Incident Form

Date Of Incident *
Mm/Dd/Yyyy

Location Observed *

Make & Model *

License Plate Number
0 Of 7 Max Characters

Company (If Known) *

Description *
Please Provide A Description Of The Incident.
0 Of 1000 Max Characters

Contact Information

If You Would Like To Provide Any Additional Feedback, Please Contact Information.

Fig. 1. Left: OL 316, DMV report form for AV collisions (page 1/3). Right: DMV form for the general public to report AV incidents.

2.3 Barriers of Sharing Safety-Critical Data

To encourage AV companies to share data beyond the mandated minimum, we need further understanding of data-sharing obstacles in three aspects. First, it is essential to explore why the known benefits of data-sharing are not sufficiently motivating. Historically, there are instances where data and knowledge sharing have created collective benefits for the autonomous industry. For example, Volvo's Nils Bohlin relinquished the original safety belt patent, allowing it to be shared freely [6]. This act of sharing has reportedly saved millions of lives.

Second, while there are known barriers to general data-sharing in the automobile industry, it remains unclear whether these barriers generalize to sharing AV safety-critical data. Some of these known barriers include lack of organizational knowledge for data sharing [23, 59], technical hurdles for moving and storing data [66], and concerns about violating legislation for privacy and user protection [43, 66, 69].

Finally, there might be unknown obstacles that discourage data-sharing practices. We set off our study to reveal such additional challenges. Without a fundamental understanding of these data-sharing barriers, it remains challenging to motivate data-sharing practices effectively. Although various solutions have been proposed, their effectiveness is often limited by a lack of comprehensive understanding of the specific obstacles and resistance within the industry.

2.4 Facilitating Safety-Critical Data Sharing

Emerging work has begun to explore possible ways to facilitate AV safety-critical data-sharing, many of which proposed novel tools to support this endeavor. Some of these technical solutions focus on streamlining the *processes* of data sharing. For instance, [12] developed federated learning models for sharing AV sensor data, while [45] proposed a blockchain protocol for sharing dashcam videos. Other prior work has devised data-sharing tools that address scale and privacy barriers of AV data [33, 50].

Some work also studied work practices in the industry to understand and identify factors that support data-sharing. Plenty of these studies emphasized the important role of human experts in enabling smooth data-sharing, as tactical knowledge was key to such processes [56, 74].

To date, few of these proposed methods have successfully promoted safety-critical data-sharing. Increasing attention has been paid to this issue within the AutoUI community. For instance, a recent

	Proactive Identification	**Reactive Discovery**	**Data Enrichment**
Reports about autonomous vehicles	In the future AV event data recorder triggered events, however data will only be released by court order [12]	AV crash reports [53, 76], police and news reports [31, 32, 42], crowd-sourced [3]	Augmenting reports with additional data (e.g. location information [73, 80])
Driven vehicle (dashcam)	Classification of manual-driven vehicle near crash events [2, 37, 64, 70, 71]	*Not* possible with monocular dashcam video	Modeling human driver behavior from manual drives [11, 44], sourcing difficult objects from video [49]
Test vehicle	*No* real world autonomous driving safety critical dataset	*No* real world in-loop autonomous driving safety database	Enriched AV dataset (e.g. extra labels or scenes [8, 10, 48])
Simulation Environments	Manual scenario design (out of domain objects, in domain objects in unique configurations, weather conditions [21, 40, 51, 62])	Anomaly in simulation (e.g. crash [5, 21] including adversarial discovery of such [20, 38, 75])	Automated augmentation from 'real' road safety critical events e.g. from dashcams [4, 21, 68]

Table 1. Publishing AV researchers source safety-critical driving data from public and open-sourced data.

workshop at the AutoUI Conference has directly called for input to address this challenge [24]. This paper contributes to this growing body of literature.

3 METHOD

To understand barriers to sharing safety-critical AV data, we interviewed twelve industry insiders who actively work on designing, developing, or researching AV safety design with large-scale data. The motivation for our methods is rooted in the rich history of Computer-Supported Cooperative Work research, which has extensively examined the work practices of professionals in software and technology development [39, 56, 60, 65]. To this day, very little research has investigated the work practices of autonomous vehicle data workers. We fill this research gap in the present study. We unfold three primary topics with our interview participants: (1) how they currently manage and work with AV safety-critical data, and the common challenges they encounter through these current practices; (2) what attitudes and rationales they hold toward sharing safety-critical AV data with the greater AV design and research community; (3) what more desirable practices of working with data they would like to propose and act toward.

Our interviews took place throughout 2023, a transformative phase in AI marked by innovations such as ChatGPT. Concurrently, the autonomous vehicle sector witnessed volatility, including layoffs at Waymo, the closure of the Argo AI venture, General Motors Cruise reducing their fleet size due to incidents and ongoing delays in Tesla's self-driving package [7, 58, 72].

3.1 Participants

We recruited participants through our extended professional networks to access insiders with substantial insights into the competitive and specialized field of autonomous driving. All participants were from different companies that were committed to developing fully autonomous vehicles. Our wide range of participants design AVs in either conventional automotive companies or specialized technology companies. Participants' demographics and professional experiences were reported in Table 2.

ID	Gender	Project	Function	Company	Exp. (years)
			Classic Automotive Industry		
1	Male	AV design	Research engineer	Vehicle manufacturer (VM)	≈ 8
4*	Male	Distraction modeling	Data-scientist & researcher	AD department of VM	≈ 6
6	Male	Vehicle perception	Research scientist	AD division of VM	≈ 7
8	Male	Driving performance	Interface researcher	Research division of VM	> 10
10	Male	Behavior modeling	Human factors researcher	Safety research division of VM	≈ 12
12	Male	Vehicle perception	Computer vision engineer	AD department of vehicle supplier	≈ 15
			Specialized Autonomous Driving (AD) Technology		
2	Male	Technical direction	Lead of AD research	AD package provider	> 10
3*	Male	Lidar motion estimation	Algorithm engineer	AD software developer	≈ 6
5	Female	Pedestrian behavior	Qualitative researcher	AD software and service developer	> 10
7	Male	Cross project alignment	Infrastructure manager	Mixed-terrain AD provider	> 15
9*	Male	AD research	Professor	Driving specialized university	≈ 8
11	Female	AD annotation	Engineering manager	AD vehicle producer	≈ 5

Table 2. Participant details. Asterisks (*) indicate academic affiliations. Company affiliations are obfuscated.

3.2 Interviews

We conducted our studies through online interviews, examining participants' current practices and exploring their data needs and barriers. The Institutional Review Board of the authors' affiliated institute reviewed and approved the study protocol. The interviews were 50–90 minutes long, semi-structured, and dependent on the type of data the practitioners were working with. The interview protocol *will be* made available as supplementary material. We guaranteed the interviewee's anonymity, and the transcripts will not be made available.

3.3 Data Analysis

We transcribed recordings of the interview sessions for data analysis. Utilizing an iterative inductive coding method [15], two main authors extracted initial codes and later used an affinity diagram to organize them into themes. This approach was chosen over grounded theory because it allows themes to emerge from the data without needing a predefined theoretical framework, enabling a more natural identification of patterns relevant to our specific context.

We next applied journey mapping [26] to trace the workflows of how these practitioners manage and work with mass-scale AV data.

We edited quotes within this paper lightly for clarity and readability. We removed speech disfluencies but did not alter the meaning or context of the participants' statements.

4 FINDINGS

Findings from the interviews suggested AV safety-critical data embedded several types of knowledge that were crucial to advancing AV safety design. As such, our participants were mostly unwilling to make such data publicly available, as they believed it should be private knowledge that yielded a competitive advantage for an AV company. While AV safety has been a targeted item for fierce competition across the AV industry, this only reinforced practitioners' preferences to keep data as internal resources. In contrast to common beliefs in prior literature, which proposed advancing tools to unblock data-sharing barriers, the risks of revealing critical AV safety design knowledge embedded in data are the root causes of practitioners' hesitations.

We structure this Findings section as follows: We first provide an overview of our participants' approaches to working with AV safety-critical data. This offers some context as we unfold our findings. We then elaborate on the key obstacles to data sharing per participants' views (i.e., risks of sharing specific knowledge of AV safety design). Finally, we further unpack their rationales behind why such data should be private knowledge.

Overview of participants' AV safety design work practices. First, most practitioners had constant access to company-owned data sources (see Table 3). While massive data continuously came in as data *streams*, they attained up-to-date data from these company-owned sources on demand, instead of working with one-time, fixed datasets. All participants predominantly worked with data collected by their own companies. Participant 2 (P2) elaborated, "Even some of those biggest companies with a lot of money, vehicles, customers, and users, they still only have their data" to work with.

Table 3 categorizes the approaches autonomous driving practitioners use to source safety-critical driving data. We group them into three categories, similarly to Table 1. Proactively looking for safety-critical data, e.g., by tuning the vehicle collection parameters remotely to send data in-house for specific scenarios (such as vehicles overtaking in tunnels); discovering safety-critical data reactively, such as annotation from a safety driver ("uncomfortable side swerve"); and enriching existing data to create more safety-critical data, such as generating variations of safety-critical confluences.

All AV data practitioners had access to a basic AV data pipeline, which generally consists of policies for data collection, targeted strategies to capture specific data, methods for organizing and filtering the amassed data, systems for developing and refining AI models, mechanisms to identify and respond to machine learning failures, and strategies for the long-term management and application of the gathered data. Specialized AD companies offer more organizational support for data work, such as implementing robust data governance frameworks, establishing dedicated teams for real-time data monitoring, providing advanced tools for data wrangling and curation, and facilitating ongoing training for AI model development and failure analysis (see Figure 2). Practitioners gradually process data through small, repeated iterative steps involving collecting, organizing, refining, reviewing, oversight, and repurposing data within a well-defined workflow for safety-critical AV data tasks.

No practitioner reported using open and external safety-critical data for their AV design. Public datasets were of low importance to most interviewees' daily AV design work, except for P3 who worked on autonomous driving algorithm improvements in affiliation with an academic institution. Instead, practitioners applied these public datasets for non-design purposes, such as gaining visibility for their work (P3, P7), facilitating training processes (P2, P8, P9), and setting benchmarks for algorithmic performance (P3, P5).

4.1 Risks of publicizing AV design knowledge are key barriers to data-sharing

All of our participants expressed hesitations about sharing AV safety-critical data. They confirmed several barriers that were recognized in the prior literature (see subsection 2.3), including concerns over privacy, the vast scale of data, and access to infrastructure. These challenges, while significant, are not unmanageable; each company has established technical solutions to mitigate these issues effectively. On top of these, practitioners mentioned a more fundamental concern: *Data inherently*

	Proactive Identification	**Reactive Discovery**	**Data Enrichment**
Deployed vehicles	- Predetermined data of interest - Remote tuning of fleet collection parameters	- Driver disengagement - Bug report - AV disengagement	- Filtering in labeled database to identify related retained data
Testing vehicles	- Driving in safety critical situations (hours, weather, location) - Staging situations (testing track, dummies)	- Supervisor disengagement - Driving style/situation feedback - Bug report - AV disengagement	- Recreating variations of similar driving situations - Reviewing all data (skimming) to identify related behavior
Simulation environment	- Manual expert-informed scenario design	- Anomaly in simulation (e.g., crash)	- Automated augmentation from 'real' road safety critical events

Table 3. Autonomous driving practitioners source safety critical driving data from three sources using three general approaches.

carried critical knowledge about how their companies designed and developed AVs. The risk of leaking such design knowledge primarily prevented them from sharing data. They believed such knowledge should be kept as an internal resource, as it was key to an AV company's competitive edge.

Participants pinpointed at least four types of AV design knowledge that were embedded in AV safety-critical data and could be revealed through data sharing; these include: (1) how their company defined and operationalized AV safety; (2) how they constructed their ML infrastructure; (3) where AVs were most susceptible to failure modes; (4) where a handover of safety liability could take place.

Participants first highlighted that there was no standard approach toward defining AV safety at scale, and thus, how a company defined and operationalized this concept was a strategic decision of its own. Likewise, there is no consensus on the causes of safety-critical events for autonomous driving. Participants suggested these ideas were often reflected in how a company sourced data; specifically, when and where data was collected indicated which scenarios were considered safety hazards by an AV design team. For instance, P12 mentioned a serious AV collision that involved a bicycle or motorcycle rider would likely result in the rider lying on the ground after being hit by the vehicle. Therefore, identifying a person lying on the ground became a key indicator that helped their team define this specific type of safety-critical event. However, such contextual knowledge might not be widely shared by all practitioners.

Likewise, participants suggested that how data was collected, stored, and annotated indicated how it would later on be used for ML development. Oftentimes, this revealed information about key parameters and data structure of ML models. Practitioners also feared that "data carries knowledge downstream". For example, P12 suggested that "[one] could get insight on how all sensors are designed based on the data." P1 further specified that data could even reveal information about models and prototypes that have not yet been released:

> "*A lot of the data is off internal systems or, like, internal prototypes that haven't ever made it to production or of, you know, internal features, internal software sets, internal sensor sets that haven't been made to production. And so, we do not want any of that information out in public either. So, none of those data sets can be shared.*" – P1

Besides, many participants mentioned that their companies created their own suite of tools to pre-process data for their ML models, and AV safety-critical data carried along knowledge about the

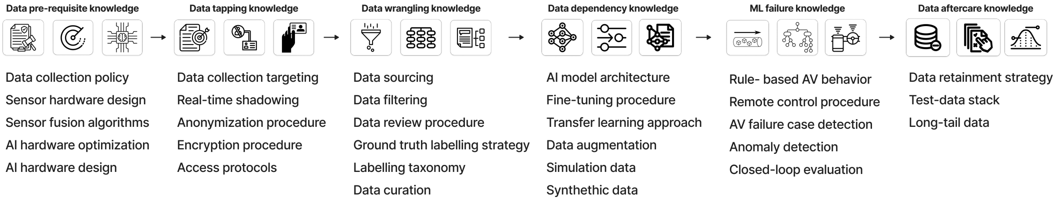


Fig. 2. AV developers are reluctant to share data, that encodes precious knowledge, across the industrialized AV data pipeline.

design of these internal tools. In P1's company, data was collected and stored in a way that allowed their internal dashboard to readily parse out the location and time where safety-critical events took place. Alternatively, all data at P11's firm would first go through a system that automatically triaged different types of AV crashes. Putting together, P8 suggested that AV companies "built these automated systems because they had consistent goals and consistent perspectives on how they want to use the data" for their AV design. On the flip side, attaining data would allow one to infer the metrics, goals, and perspectives held internally at an AV company.

According to our participants, AV safety-critical data not only hinted at what constituted their AV design, but it also gave clues to how their design might break. Participants mentioned identifying bugs in their ML models was a crucial step in improving the safety performance of AVs. As such, what their AV safety-critical datasets entailed informed where their models were subject to malperformance at the moment, how they intended to debug such issues, and eventually, how they advanced AV safety.

> "*But then we need to identify what the issues are. And then a big part of my job was kind of going through all the images that the model failed on. And so if it failed on it, you had to go and triage why did it fail. And if there was a reason for why something failed, you would then go in and add more images into the training set for that specific scenario. And then redo the whole process, and then go over it again to see if the model failed on that scenario again. And if it didn't, then that meant that you fixed that issue. And then you had to go through for the next issue and stuff like that. So a long process of iterating back and forth between evaluation sets.*" – P11

Finally, participants mentioned that safety-critical data could also entail information about legal liability. On one hand, owning more data placed more responsibility on AV companies to improve the safety of their vehicles. On the other hand, "a lot of these rare events are technically illegal actions by drivers" (P1). Many safety-critical incidents took place when drivers were "not paying attention on the road" or "on their phones when they were driving." Therefore, participants acknowledged that making such data available would raise complex questions about who should be held accountable for AV safety-critical events. Consequently, drivers might backfire and become unwilling to grant permission to collect their data.

4.2 Concerns about sharing AV safety-critical data as public knowledge

4.2.1 Too insightful to share: Data work produces intellectual property . Participants believed such embedded knowledge in AV safety-critical data should be treated as private assets. While they acknowledged on-road safety as a public good, the process of handling data reshaped it as intellectual property and a competitive advantage for AV companies, which caused reluctance for data-sharing.

How did AV companies establish competitive edges through the ways they worked with safety-critical data? Our participants indicated three common means:

(1) *Remarking areas in need of new design solutions.* Many of our participants admitted pinning down design problems for AV safety improvement could often be more time- and resource-consuming than generating the solutions per se. As P2 and others elaborated, a wide variety of factors could cause hazardous driving conditions, ranging from light and environmental conditions, unusual objects and animals, and atypical traffic to individual pedestrians' and drivers' behaviors. Identifying on-point design problems required understanding the full picture of driving scenarios. The types of data later on used for safety-critical design summarized AV designers' and developers' insights into the primary causes of safety-critical events.

> "*So, this is basically raw data. There's no labeling, for example, [for] a critical incident. There's no labeling that there was a takeover failure. So, for every analysis that is being made on top of the raw data, you have to define yourself. If you want to know, for example, triggers of a takeover request, then you would need to look into data for these and then you would need to define your labels in that case. There's no annotation that is somehow magically annotating videos.*" – P4

(2) *Identifying safety-critical events from mass data streams.* Participants suggested that even identifying safety-critical events *per se* from vast data streams is a highly resource- and expertise-demanding effort. As P7 elaborated, there was no handy way to "look for the top 20 scenarios when [a] vehicle struggled." This is because AV data not only came in with massive volume but also widely differing data structures. Therefore, having a "database that is searchable with a fairly wide set of parameters" would be a tremendous help for AV designers and developers to identify critical events among mass data.

Indeed, most organizations used specialized tools to help with these onerous tasks, such as searching and querying scenarios, labels, and conditions related to safety-critical events. However, even creating "a tool that allows you to query across all those different diverse data and data protocols, considering past and future data, is challenging" (P7). Furthermore, P2 suggested the design of each tool embedded "lessons learned when you encountered a specific AV safety problem." Many of our participants' companies had dedicated teams to work on creating effective tools with vast data streams, while most spent a substantial amount of their budgets to procure external tools, hire expert contractors, or even acquire specialized start-ups. Although several participants indicated their companies held the ultimate goal to "automate all manual steps in between" data work (P11), they also admitted the job of identifying safety-critical events will remain highly labor-intensive in the foreseeable future.

(3) *Herding unrevealed AV design insights.* Participants also acknowledged insights from data did not always become apparent in the first place. The fact that practitioners took a long time to gauge the value of a dataset was reflected in one of their common work practices – they seldom absorbed data at once and constantly revisited and wrangled with data. Before fully figuring out the utility of their data for safety design, most practitioners would hold up with their data given its potential competitive edges or would be advised by their stakeholders (e.g., OEMs) to do so.

4.2.2 Too competitive to share: Competitive landscape in the AV industry. Participants suggested that fierce competition across the AV industry reinforced the importance of establishing competitive edges through safety-critical data. Each believed their company-owned data gave the unique competitive advantages for them to tackle AV safety design challenges and generate one-of-a-kind solutions.

> "*The data is the new gold, because using the data, you can develop the solutions. And if you're the first to develop the solutions, it means you're the first to go to market.*" – P2

> "*Everybody thinks they have a unique competitive advantage. I guess everybody is hoping they have the only dataset that has all the secret information. Those data are considered valuable. And everybody is not willing to share them.*" – P12

Many participants believed their organizations held leading positions in specific types of data work knowledge, although they also acknowledged no company was advantageous in all aspects. As P2 elaborated, "each [autonomous] vehicle out there is contributing to solving a particular [AV design] problem." Hence, each AV company could claim unique competitive advantages through specific types of data work knowledge, such as: a unique data collection policy (P2, P11), data refinement strategies on hardware and software level (P2, P4), unique tooling for re-targeting data collection targets in the fleet (P8, P11), leading anonymization and data encryption procedures (P2), data sourcing and filtering tools (P5, P11, P12), leading data labeling procedures and taxonomies (P5), safe rule-based AV algorithms (P3, P7), secret AI architecture (P5, P11), fast closed-loop evaluation methods (P7), transfer learning to scale between countries and cities (P11), field-tested synthetic data augmentation approaches (P7), or knowledge about dealing with the scale of data unlike other industries (P6, P11). In summary, a lot of AV companies felt they were ahead of the competition with some crucial safety-critical data knowledge (See Figure 2).

In particular, AV safety has typically been viewed as the frontier of AV design innovation, and being able to act upon safety-critical data needs rapidly has become the key competing ground, as the expensive organizational decisions outlined by one of our interviewees suggested:

> "*Initially they were outsourcing the labeling. But then there was a big push of keeping the data in-house and also investing in an in-house data team. We had our whole called data annotation org, which was 600 700 people, All of the team was in the US, for quality, speed and safety reasons.*" – P11

However, even under this competitive landscape, participants still saw the possibilities and advantages of collaborations. They coined the term "*untrusted collaborations*" and believed these types of strategic partnerships would more likely take place when data-sharing was not a prerequisite. For example, teams at P2's company worked on building federated learning models that allowed different organizations to share lessons learned from data *without* directly sharing their safety-critical data. They elaborated with further details:

> " *Essentially, it's a way to allow learning from data but in secure enclaves, so that you're not essentially stealing the data or sharing the data. But what you're doing is [...] sharing a common solution, which is maybe a model. And then I'm going to enable that model to be training my data. And then I'm going to benefit from that, that model being improved on my data without you having to share the data.* " – P2

Participants mentioned similar approaches have been adopted in sharing the healthcare domain, leading to significant, "30 to 40% improvements on the performance of these algorithms by learning from these siloed datasets." This painted a more plausible future for practitioners to collectively contribute to advancing AV safety design.

5 DISCUSSION

Findings from the present study show practitioners' reluctance to share safety-critical data is not merely a technical or procedural issue; it is indicative of a deeper need for a paradigm shift. Prior work has concentrated on developing tools to facilitate data sharing, yet our findings indicate a different challenge: safety-critical data inherently embeds specific AV safety design knowledge. This nature of data causes fundamental obstacles to data sharing, as practitioners feared sharing data would share specific knowledge about AV design knowledge. The perception of safety-critical data as a competitive advantage – instead of public knowledge – held practitioners back from sharing data that could advance AV safety design.

Based on these key findings, we discuss new approaches to motivating data-sharing practices in the AV design field. We recommend alternative directions for technical solutions, AV safety assessment, and legislative interventions. Furthermore, we summarize key takeaways for AutoUI researchers and suggest actionable items that they can adopt in their research processes going forward. Together, we envision a future where the competitive landscape of the AV industry can be leveraged for collective goods.

5.1 Proposed Approaches to Motivating Data-Sharing Practices

Grounded on our findings, we suggest plausible ways to motivate safety-critical data-sharing. Specifically, we build on existing efforts (i.e., technical solutions, work practices, and relevant policies) that address this issue and propose new directions for each. We recommend new approaches to (1) debating and stratifying public and private AV safety knowledge, (2) innovating data tools and data sharing pipelines that enable easier sharing of public AV safety data *and knowledge*; (3) offsetting costs of curating safety-critical data and incentivizing data sharing.

Protecting data-sharing from knowledge-sharing. We first suggest that researchers and developers conceive technical solutions that enable sharing data without sharing design knowledge, given that abundant prior work has already been dedicated to developing novel data-sharing tools that address scale and privacy concerns [33, 50]. Our findings suggest that prior efforts might not directly address practitioners' concerns (i.e., possible leaks of AV design knowledge), as these proposed tools mostly target smoothing the process of data sharing. Instead, developers of these tools should consider means to separate AV safety-critical data from its embedded knowledge.

Assessing AV safety without accessing data. While limited data-sharing will likely continue in the near future, advancing AV safety cannot wait. We ask whether there can be alternative approaches to leveraging safety-critical data from AV companies without having them directly share their data. As existing work has made many attempts to simulate environments and scenarios for AV development and testing [22, 34, 41, 77], we encourage practitioners to jointly contribute to building standardized simulated platforms for AV safety assessment. This might allow practitioners to apply insights they acquire from AV safety-critical data without sharing data first-hand.

Incentivizing data-sharing through federated programs. Besides input from AV practitioners, official mandates can accelerate and make many of the aforementioned proposals more effective. Advances in other technology fields have shown great promises of policy interventions [67, 78]. In particular, legislation can offer unique help by developing and launching incentive programs to share data, as newly proposed policies, like the EU Data Act, allow ways out for data considered trade secrets [27]. We learn from our interviewees that the lack of incentives remains a major hurdle for data-sharing, as it typically takes abundant resources to collect safety-critical data. Executing incentive programs that can help offset such costs will likely provide immediate motivation for data-sharing.

5.2 Limitations

Although we aimed for a diverse sample of industry practitioners in autonomous vehicle design, this interview study could be skewed by the interviewees we were able to engage. While all participants were recruited from different companies, we relied on personal networks for referrals, as open sampling proved difficult due to companies' interest in protecting intellectual property. This might bias the findings to a group of participants who are more open to discussing proprietary practices. These interviews also present a Western viewpoint and exclude processes and attitudes in the

East and Global South. Lastly, the study relied on interviews and could benefit from ethnographic observation of data work practices to cross-verifying the accuracy of the practices and perceptions shared by the participants.

6 CONCLUSION

Our study reveals significant barriers that prevent autonomous vehicle (AV) companies from sharing safety-critical data, despite the clear benefits of such sharing for advancing AV safety. Through interviews with industry insiders, we identified two primary obstacles: the inherent embedding of critical knowledge within the data and the perception of safety knowledge as a competitive asset rather than a public good.

These findings highlight the need for a paradigm shift in how data sharing is approached. Rather than focusing solely on technical solutions to facilitate data exchange, it is essential to address the underlying incentives and strategic concerns of AV companies. We propose concerted efforts from academics, policymakers, and industry practitioners to create technical solutions, policy interventions, and collaborative frameworks to mitigate these barriers.

ACKNOWLEDGMENTS

We express our gratitude to current and former employees at the following automated driving companies and research centers: [Anonymized Institutions]. Their willingness to participate in interviews or guide to suitable candidates has been invaluable. We are also indebted to numerous researchers who have offered insights on the early drafts of this work.

This research has been generously supported by [Anonymized].

REFERENCES

[1] Aaron H. Jacoby, Jon S. Bouker, and Gordon Sung. 2023. NHTSA Urged to Extend Oversight of Autonomous Vehicles. https://www.natlawreview.com/article/us-labor-groups-seek-greater-dot-and-nhtsa-oversight-autonomous-vehicles

[2] Ramin Arvin, Asad J. Khattak, and Hairong Qi. 2021. Safety critical event prediction through unified analysis of driver and vehicle volatilities: Application of deep learning methods. *Accident Analysis and Prevention* 151 (March 2021), 105949. https://doi.org/10.1016/j.aap.2020.105949

[3] Elon Bachman and Ian Capulet. 2023. Tesla Deaths: Digital record of Tesla crashes resulting in death. https://doi.org/10.7910/DVN/MCNENT

[4] Sai Krishna Bashetty, Heni Ben Amor, and Georgios Fainekos. 2020. DeepCrashTest: Turning Dashcam Videos into Virtual Crash Tests for Automated Driving Systems. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 11353–11360. https://doi.org/10.1109/ICRA40945.2020.9197053 ISSN: 2577-087X.

[5] Joe Beck, Ramin Arvin, Steve Lee, Asad Khattak, and Subhadeep Chakraborty. 2023. Automated vehicle data pipeline for accident reconstruction: New insights from LiDAR, camera, and radar data. *Accident Analysis & Prevention* 180 (Feb. 2023), 106923. https://doi.org/10.1016/j.aap.2022.106923

[6] Douglas Bell. 2019. Volvo's gift to the world, modern seat belts have saved millions of lives. https://www.forbes.com/sites/douglasbell/2019/08/13/60-years-of-seatbelts-volvos-great-gift-to-the-world/ Section: Entrepreneurs.

[7] Rebecca Bellan. 2023. Waymo cuts 200 employees after second round of layoffs. https://techcrunch.com/2023/03/01/waymo-cuts-200-employees-after-second-round-of-layoffs/

[8] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. 2021. The Fishyscapes Benchmark: Measuring Blind Spots in Semantic Segmentation. *International Journal of Computer Vision* 129, 11 (Nov. 2021), 3119–3135. https://doi.org/10.1007/s11263-021-01511-6

[9] Daniel Bogdoll, Felix Schreyer, and J. Marius Zöllner. 2023. Ad-datasets: A Meta-collection of Data Sets for Autonomous Driving. 46–56. https://www.scitepress.org/Link.aspx?doi=10.5220/0011001900003191

[10] Daniel Bogdoll, Svenja Uhlemeyer, Kamil Kowol, and J. Marius Zöllner. 2023. Perception Datasets for Anomaly Detection in Autonomous Driving: A Survey. In *2023 IEEE Intelligent Vehicles Symposium (IV)*. 1–8. https://doi.org/10.1109/IV55152.2023.10186609 arXiv:2302.02790 [cs].

[11] Jonas Bärgman, Christian-Nils Boda, and Marco Dozza. 2017. Counterfactual simulations applied to SHRP2 crashes: The effect of driver behavior models on safety benefit estimations of intelligent safety systems. *Accident Analysis &*

Prevention 102 (May 2017), 165–180. https://doi.org/10.1016/j.aap.2017.03.003

[12] Klaus Böhm, Tibor Kubjatko, Daniel Paula, and Hans-Georg Schweiger. 2020. New developments on EDR (Event Data Recorder) for automated vehicles. *Open Engineering* 10, 1 (Jan. 2020), 140–146. https://doi.org/10.1515/eng-2020-0007 Publisher: De Gruyter Open Access.

[13] California Department of Motor Vehicles. 2020. OL 316, Report of Traffic Collision Involving an Autonomous Vehicle. https://www.dmv.ca.gov/portal/file/cruise_083021/

[14] California Department of Motor Vehicles. 2024. Autonomous Vehicles Incident Form - California DMV. https://www.dmv.ca.gov/portal/dmv-autonomous-vehicles-feedback-form/

[15] Sheelagh Carpendale, Søren Knudsen, Alice Thudt, and Uta Hinrichs. 2017. Analyzing Qualitative Data. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces (ISS '17)*. Association for Computing Machinery, New York, NY, USA, 477–481. https://doi.org/10.1145/3132272.3135087

[16] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. 2023. End-to-end Autonomous Driving: Challenges and Frontiers. (2023). https://doi.org/10.48550/ARXIV.2306.16927 Publisher: [object Object] Version Number: 1.

[17] Dan Luu. 2024. Notes on Cruise's pedestrian accident. https://danluu.com/cruise-report/

[18] Daniel Atherton. 2023. Incident 596: Cruise's Autonomous Vehicles Allegedly Engaging in Risky Behavior Near Pedestrians. https://incidentdatabase.ai/cite/596/

[19] J. C. F. de Winter, D. Dodou, R. Happee, and Y. B. Eisma. 2019. Will vehicle data be shared to address the how, where, and who of traffic accidents? *European Journal of Futures Research* 7, 1 (March 2019), 2. https://doi.org/10.1186/s40309-019-0154-3

[20] Wenhao Ding, Baiming Chen, Minjun Xu, and Ding Zhao. 2020. Learning to Collide: An Adaptive Safety-Critical Scenarios Generating Method. http://arxiv.org/abs/2003.01197 arXiv:2003.01197 [cs].

[21] Wenhao Ding, Chejian Xu, Mansur Arief, Haohong Lin, Bo Li, and Ding Zhao. 2023. A Survey on Safety-Critical Driving Scenario Generation - A Methodological Perspective. *IEEE Transactions on Intelligent Transportation Systems* 24, 7 (July 2023), 6971–6988. https://doi.org/10.1109/TITS.2023.3259322

[22] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*. PMLR, 1–16. https://proceedings.mlr.press/v78/dosovitskiy17a.html ISSN: 2640-3498.

[23] Christian Dremel. 2017. Barriers to the adoption of big data analytics in the automotive sector. *AMCIS 2017 Proceedings* (Aug. 2017). https://aisel.aisnet.org/amcis2017/AdoptionIT/Presentations/11

[24] Patrick Ebel, Pavlo Bazilinskyy, Angel Hsing-Chi Hwang, Wendy Ju, Hauke Sandhaus, Aravinda Ramakrishnan Srinivasan, Qian Yang, and Philipp Wintersberger. 2023. Breaking Barriers: Workshop on Open Data Practices in AutoUI Research. In *Adjunct Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 227–230. https://doi.org/10.1145/3581961.3609835

[25] Patrick Ebel, Florian Brokhausen, and Andreas Vogelsang. 2020. The Role and Potentials of Field User Interaction Data in the Automotive UX Development Lifecycle: An Industry Perspective. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Virtual Event DC USA, 141–150. https://doi.org/10.1145/3409120.3410638

[26] Anja Endmann and Daniela Keßner. 2016. User Journey Mapping – A Method in User Experience Design. *i-com* 15, 1 (April 2016), 105–110. https://doi.org/10.1515/icom-2016-0010 Publisher: Oldenbourg Wissenschaftsverlag.

[27] European Commission. 2024. Data Act explained. https://digital-strategy.ec.europa.eu/en/factpages/data-act-explained. Accessed: 2024-7-2.

[28] European Commission - Have your say. 2021. Vehicle safety – technical requirements & test procedures for EU type-approval of event data recorders (EDRs). https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12989-Vehicle-safety-technical-requirements-test-procedures-for-EU-type-approval-of-event-data-recorders-EDRs-/feedback_en?p_id=26647797

[29] European Transport Safety Council. 2022. Car black boxes will be virtually useless to safety researchers – ETSC. https://etsc.eu/car-black-boxes-will-be-virtually-useless-to-safety-researchers/

[30] Mark Fagan. 2023. A Brief for Policymakers on the Regulatory Landscape for Autonomous Vehicles Data Sharing and Privacy. *Harvard Kennedy School Policy Brief* (2023).

[31] Alessio Gambi, Tri Huynh, and Gordon Fraser. 2019. Automatically Reconstructing Car Crashes from Police Reports for Testing Self-Driving Cars. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*. 290–291. https://doi.org/10.1109/ICSE-Companion.2019.00119 ISSN: 2574-1934.

[32] Alessio Gambi, Tri Huynh, and Gordon Fraser. 2019. Generating effective test cases for self-driving cars from police reports. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2019)*. Association for Computing Machinery, New York, NY,

USA, 257–267. https://doi.org/10.1145/3338906.3338942

[33] Gonzalo Munilla Garrido, Kaja Schmidt, Christopher Harth-Kitzerow, Johannes Klepsch, Andre Luckow, and Florian Matthes. 2021. Exploring privacy-enhancing technologies in the automotive value chain. *2021 IEEE International Conference on Big Data (Big Data)* (Dec. 2021), 1265–1272. https://doi.org/10.1109/BigData52589.2021.9671528 Conference Name: 2021 IEEE International Conference on Big Data (Big Data) ISBN: 9781665439022 Place: Orlando, FL, USA Publisher: IEEE.

[34] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, John D. Co-Reyes, Rishabh Agarwal, Rebecca Roelofs, Yao Lu, Nico Montali, Paul Mougin, Zoey Yang, Brandyn White, Aleksandra Faust, Rowan McAllister, Dragomir Anguelov, and Benjamin Sapp. 2023. Waymax: An Accelerated, Data-Driven Simulator for Large-Scale Autonomous Driving Research. https://doi.org/10.48550/arXiv.2310.08710 arXiv:2310.08710 [cs].

[35] Junyao Guo, Unmesh Kurup, and Mohak Shah. 2018. Is it Safe to Drive? An Overview of Factors, Challenges, and Datasets for Driveability Assessment in Autonomous Driving. https://doi.org/10.48550/arXiv.1811.11277 arXiv:1811.11277 [cs].

[36] Xiaoyu Guo and Yunlong Zhang. 2022. Maturity in Automated Driving on Public Roads: A Review of the Six-Year Autonomous Vehicle Tester Program. *Transportation Research Record: Journal of the Transportation Research Board* 2676 (June 2022), 036119812210927. https://doi.org/10.1177/03611981221092720

[37] Jonathan M Hankey, Miguel A Perez, and Julie A McClafferty. 2016. *Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets.* Technical Report. Virginia Tech Transportation Institute.

[38] Niklas Hanselmann, Katrin Renz, Kashyap Chitta, Apratim Bhattacharyya, and Andreas Geiger. 2022. KING: Generating Safety-Critical Driving Scenarios for Robust Imitation via Kinematics Gradients. (2022). https://doi.org/10.48550/ARXIV.2204.13683 Publisher: [object Object] Version Number: 1.

[39] Amy K. Heger, Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–29. https://doi.org/10.1145/3555760

[40] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. 2022. Scaling Out-of-Distribution Detection for Real-World Settings. http://arxiv.org/abs/1911.11132 arXiv:1911.11132 [cs].

[41] Philipp Hock, Johannes Kraus, Franziska Babel, Marcel Walch, Enrico Rukzio, and Martin Baumann. 2018. How to Design Valid Simulator Studies for Investigating User Experience in Automated Driving: Review and Hands-On Considerations. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18)*. Association for Computing Machinery, New York, NY, USA, 105–117. https://doi.org/10.1145/3239060.3239066

[42] James C. Holland and Arman Sargolzaei. 2020. Verification of Autonomous Vehicles: Scenario Generation based on Real World Accidents. In *2020 SoutheastCon*, Vol. 2. 1–7. https://doi.org/10.1109/SoutheastCon44009.2020.9368284 ISSN: 1558-058X.

[43] Mingfu Huang, Rushit Dave, Nyle Siddiqui, and Naeem Seliya. 2021. Examining Modern Data Security and Privacy Protocols in Autonomous Vehicles. *International Journal of Computer Science and Information Technology* 13, 5 (Oct. 2021), 01–19. https://doi.org/10.5121/ijcsit.2021.13501

[44] Amany A. Kandeel, Ahmed A. Elbery, Hazem M. Abbas, and Hossam S. Hassanein. 2021. Driver Distraction Impact on Road Safety: A Data-driven Simulation Approach. In *2021 IEEE Global Communications Conference (GLOBECOM)*. 1–6. https://doi.org/10.1109/GLOBECOM46510.2021.9685932

[45] Keonhyeong Kim and Im Y. Jung. 2020. Encouraging data sharing for safe autonomous driving. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 1–5. https://doi.org/10.1109/PerComWorkshops48775.2020.9156209

[46] Philip Koopman and Michael Wagner. 2016. Challenges in Autonomous Vehicle Testing and Validation. *SAE International Journal of Transportation Safety* 4, 1 (April 2016), 15–24. https://doi.org/10.4271/2016-01-0128

[47] Timothy B. Lee. 2023. Are self-driving cars already safer than human drivers? https://arstechnica.com/cars/2023/09/are-self-driving-cars-already-safer-than-human-drivers/

[48] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, Xiaodan Liang, Zhenguo Li, and Hang Xu. 2022. CODA: A Real-World Road Corner Case Dataset for Object Detection in Autonomous Driving. https://doi.org/10.48550/arXiv.2203.07724 arXiv:2203.07724 [cs].

[49] Krzysztof Lis, Krishna Kanth Nakka, Pascal Fua, and Mathieu Salzmann. 2019. Detecting the Unexpected via Image Resynthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea (South), 2152–2161. https://doi.org/10.1109/ICCV.2019.00224

[50] Andre Luckow, Ken Kennedy, Fabian Manhardt, Emil Djerekarov, Bennie Vorster, and Amy Apon. 2015. Automotive big data: Applications, workloads and infrastructures. In *2015 IEEE International Conference on Big Data (Big Data)*.

1201–1210. https://doi.org/10.1109/BigData.2015.7363874

[51] Kira Maag, Robin Chan, Svenja Uhlemeyer, Kamil Kowol, and Hanno Gottschalk. 2023. Two Video Data Sets for Tracking and Retrieval of Out of Distribution Objects. In *Computer Vision – ACCV 2022*, Lei Wang, Juergen Gall, Tat-Jun Chin, Imari Sato, and Rama Chellappa (Eds.). Vol. 13845. Springer Nature Switzerland, Cham, 476–494. https://doi.org/10.1007/978-3-031-26348-4_28 Series Title: Lecture Notes in Computer Science.

[52] Carl Macrae. 2022. Learning from the Failure of Autonomous and Intelligent Systems: Accidents, Safety, and Sociotechnical Sources of Risk. *Risk Analysis* 42, 9 (2022), 1999–2025. https://doi.org/10.1111/risa.13850 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.13850.

[53] Roger McCarthy. 2021. Autonomous Vehicle (AV) Accident Data Analysis: California OL 316 Reports: 2015-2020. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg* 8 (July 2021). https://doi.org/10.1115/1.4051779

[54] Alexander G. Mirnig, Rod McCall, Alexander Meschtscherjakov, and Manfred Tscheligi. 2019. The Insurer's Paradox: About Liability, the Need for Accident Data, and Legal Hurdles for Automated Driving. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Utrecht Netherlands, 113–122. https://doi.org/10.1145/3342197.3344540

[55] Umberto Montanaro, Shilp Dixit, Saber Fallah, Mehrdad Dianati, Alan Stevens, David Oxtoby, and Alexandros Mouzakitis. 2019. Towards connected autonomous driving: review of use-cases. *Vehicle System Dynamics* 57, 6 (June 2019), 779–814. https://doi.org/10.1080/00423114.2018.1492142 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00423114.2018.1492142.

[56] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How data science workers work with data: discovery, capture, curation, design, creation *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300356

[57] Sina Nordhoff and Joost de Winter. 2023. *Why do drivers and automation disengage the automation? Results from a study among Tesla users.* https://doi.org/10.13140/RG.2.2.22115.84003

[58] Alan Ohnsman. [n. d.]. Argo AI, Ford's Self-Driving Venture With Volkswagen, Is Shutting Down. https://www.forbes.com/sites/alanohnsman/2022/10/26/argo-ai-fords-self-driving-venture-with-volkswagen-is-shutting-down/ Section: Transportation.

[59] Farhad Panahifar, Sajjad Shokouhyar, and Sina Mosafer. 2022. Identifying and assessing barriers to information sharing in supply chain - a case study of the automotive industry. *International Journal of Business Information Systems* 41, 2 (Jan. 2022), 258–288. https://doi.org/10.1504/ijbis.2022.126131

[60] Samir Passi and Steven J. Jackson. 2018. Trust in data science: collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–28. https://doi.org/10.1145/3274405

[61] Zaydoun Yahya Rawashdeh and Zheng Wang. 2018. Collaborative automated driving: a machine learning-based method to enhance the accuracy of shared information. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. 3961–3966. https://doi.org/10.1109/ITSC.2018.8569832 ISSN: 2153-0017.

[62] Stefan Riedmaier, Thomas Ponn, Dieter Ludwig, Bernhard Schick, and Frank Diermeyer. 2020. Survey on Scenario-Based Safety Assessment of Automated Vehicles. *IEEE Access* 8 (2020), 87456–87477. https://doi.org/10.1109/ACCESS.2020.2993730

[63] Andreas Riener and Johann Reder. 2014. Collective data sharing to improve on driving efficiency and safety. In *Adjunct Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Seattle WA USA, 1–6. https://doi.org/10.1145/2667239.2667266

[64] Arash Rocky, Qingming Jonathan Wu, and Wandong Zhang. 2024. Review of Accident Detection Methods Using Dashcam Videos for Autonomous Driving Vehicles. *IEEE Transactions on Intelligent Transportation Systems* (2024), 1–19. https://doi.org/10.1109/TITS.2024.3354852 Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[65] Annabel Rothschild, Amanda Meng, Carl DiSalvo, Britney Johnson, Ben Rydal Shapiro, and Betsy DiSalvo. 2022. Interrogating data work as a community of practice. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–28. https://doi.org/10.1145/3555198

[66] Daniel L. Rubinfeld and Michal Gal. 2016. Access barriers to big data. https://doi.org/10.2139/ssrn.2830586

[67] Hauke Sandhaus, Wendy Ju, and Qian Yang. 2023. Towards prototyping driverless vehicle behaviors, city design, and policies simultaneously. https://doi.org/10.48550/arXiv.2304.06639 arXiv:2304.06639 [cs].

[68] John M. Scanlon, Kristofer D. Kusano, Tom Daniel, Christopher Alderson, Alexander Ogle, and Trent Victor. 2021. Waymo simulated driving behavior in reconstructed fatal crashes within an autonomous vehicle operating domain. *Accident; Analysis and Prevention* 163 (Dec. 2021), 106454. https://doi.org/10.1016/j.aap.2021.106454

[69] Morgan Klaus Scheuerman, Katy Weathington, Tarun Mugunthan, Emily Denton, and Casey Fiesler. 2023. From Human to Data to Dataset: Mapping the Traceability of Human Subjects in Computer Vision Datasets. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 55:1–55:33. https://doi.org/10.1145/3579488

[70] Thomas Seacrist, Ethan C Douglas, Chloe Hannan, Rachel Rogers, Aditya Belwadi, and Helen Loeb. 2020. Near crash characteristics among risky drivers using the SHRP2 naturalistic driving study. *Journal of safety research* 73 (2020), 263–269.

[71] Thomas Seacrist, Ethan C Douglas, Elaine Huang, James Megariotis, Abhiti Prabahar, Abyaad Kashem, Ayya Elzarka, Leora Haber, Taryn MacKinney, and Helen Loeb. 2018. Analysis of near crashes among teen, young adult, and experienced adult drivers using the SHRP2 naturalistic driving study. *Traffic injury prevention* 19, sup1 (2018), S89–S96.

[72] Umar Shakir. 2023. Tesla pauses new Full Self-Driving beta installations until recall is addressed. https://www.theverge.com/2023/2/27/23616772/tesla-full-self-driving-fsd-beta-paused-ota-update-nhtsa-recall

[73] Amolika Sinha, Sai Chand, Vincent Vu, Huang Chen, and Vinayak Dixit. 2021. Crash and disengagement data of autonomous vehicles on public roads in California. *Scientific Data* 8, 1 (Nov. 2021), 298. https://doi.org/10.1038/s41597-021-01083-7

[74] Divy Thakkar, Azra Ismail, Pratyush Kumar, Alex Hanna, Nithya Sambasivan, and Neha Kumar. 2022. When is machine learning data good?: valuing in public health datafication. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–16. https://doi.org/10.1145/3491102.3501868

[75] Jingkang Wang, Ava Pun, James Tu, Sivabalan Manivasagam, Abbas Sadat, Sergio Casas, Mengye Ren, and Raquel Urtasun. 2021. AdvSim: Generating Safety-Critical Scenarios for Self-Driving Vehicles. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Nashville, TN, USA, 9904–9913. https://doi.org/10.1109/CVPR46437.2021.00978

[76] Song Wang and Zhixia Li. 2019. Exploring the mechanism of crashes with automated vehicles using statistical modeling approaches. *PLOS ONE* 14, 3 (March 2019), e0214550. https://doi.org/10.1371/journal.pone.0214550

[77] Chejian Xu, Wenhao Ding, Weijie Lyu, Zuxin Liu, Shuai Wang, Yihan He, Hanjiang Hu, Ding Zhao, and Bo Li. 2022. SafeBench: A Benchmarking Platform for Safety Evaluation of Autonomous Vehicles. (2022). https://doi.org/10.48550/ARXIV.2206.09682 Publisher: [object Object] Version Number: 4.

[78] Qian Yang. 2023. Designing technology and policy simultaneously: towards a research agenda and new practice | extended abstracts of the 2023 CHI conference on human factors in computing systems. https://dl.acm.org/doi/10.1145/3544549.3573827

[79] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. 2020. A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE Access* 8 (2020), 58443–58469. https://doi.org/10.1109/ACCESS.2020.2983149 Conference Name: IEEE Access.

[80] Ou Zheng, Mohamed Abdel-Aty, Zijin Wang, Shengxuan Ding, Dongdong Wang, and Yuxuan Huang. 2023. AVOID: Autonomous Vehicle Operation Incident Dataset Across the Globe. (2023). https://doi.org/10.13140/RG.2.2.21627.59680 Publisher: [object Object].

[81] Siying Zhu and Qiang Meng. 2022. What can we learn from autonomous vehicle collision data on crash severity? A cost-sensitive CART approach. *Accident Analysis and Prevention* 174 (Sept. 2022), 106769. https://doi.org/10.1016/j.aap.2022.106769