



Terms and Conditions and Standard User License Agreement - Download Documents

Copyright: This Document is copyrighted. No rights therein are granted except as set forth in this License. Any copying, transmission, modification or reproduction of the copyrighted material, in part or in whole, except as herein permitted is cause for revocation of this License.

License: SAE International hereby grants you a nonexclusive, nontransferable right to download this document in an electronic format for your individual use on one computer. One copy of the document can be printed for individual use. The document may not be copied in any form for distribution to other users or other computers.

NOTICE: SAE International prohibits the entry of its Standards and related Documents into any form of Artificial Intelligence (AI) tools and further prohibits creating derivatives of such SAE content using AI without express written permission from SAE International.

General: Documents that have been successfully downloaded cannot be returned for refund or credit. This Agreement is the complete and exclusive statement of the agreement between you and SAE International and supersedes any and all prior agreements or understandings, either written or oral, concerning the subject of this Agreement. Any modifications must be in writing and signed by the parties.

Notice to Resellers: Authorized resellers of SAE International documents may download documents on behalf of their customers and forward them directly, unopened, to their customers. Resellers may not otherwise copy, transmit, modify or reproduce documents they download from SAE International.

This License shall terminate upon violation of any of its terms.

YOU ACKNOWLEDGE THAT YOU HAVE READ THIS LICENSE AGREEMENT, UNDERSTAND IT AND AGREE TO BE BOUND BY THE TERMS AND CONDITIONS THEREOF.

SAE has provided view-only access to the IWG on AI, UN WP.29 for review/reference purposes only. This document is SAE copyrighted intellectual property. It may not be shared, downloaded, duplicated, reprinted, or transmitted in any manner without prior written permission from SAE. SAE requires that you make best efforts to secure and protect the document from disclosure, taking at least the same care that you would for your own confidential information. Thank you.



SURFACE VEHICLE INFORMATION REPORT

J3298™

JUL2024

Issued

2024-07

Artificial Intelligence Data for Ground Vehicle Applications

RATIONALE

Given the rapid interest in Artificial Intelligence (AI) and the deployment of AI solutions, it is important to understand the various aspects of data associated with AI for ground mobility, to explicitly state the current issues and note the potential future issues. For the sake of industry standardization of engineering terms, definitions, and applications, the SAE Ground Vehicle Artificial Intelligence Data Task Force has prepared this document for the public, industry, and government to understand the current issues related to data for AI, including the meanings and/or limitations, to mitigate confusion.

INTRODUCTION

In 1956, John McCarthy, American computer scientist and cognitive scientist, coined the term “Artificial Intelligence,” and he defined AI as “the science and engineering of making intelligent machines.” Intelligent machines can be described as computers or computer-controlled systems that perform tasks typically associated with intelligent beings, such as humans. Today, AI is an ever-evolving, fast-moving field that is being applied to various industries. In health care, AI is being used to assist medical professionals in the detection of health-related issues via image analysis. In retail and e-commerce, AI is used by retailers to find patterns in consumer behavior to increase competitive advantage. In manufacturing, AI is being deployed to improve efficiency in the overall workflow.

In the ground vehicle domain, AI is being incorporated into the various facets of mobility, playing a key role in revolutionizing transportation and enhancing overall vehicle capabilities. AI technologies are integrated into various aspects of ground vehicles, ranging from advanced driver assistance systems (ADAS) to automated vehicles (AV). AI enables vehicles to perceive their surroundings through sensors like cameras, lidar, and radar, allowing for real-time analysis and interpretation of complex environments. Machine learning algorithms empower vehicles to make intelligent decisions based on this data, enhancing safety, efficiency, and overall performance. Additionally, AI can contribute to predictive maintenance, optimizing vehicle health and reducing downtime. The continuous evolution of AI in ground vehicles holds the promise of safer, more efficient transportation systems, with the potential to transform how we perceive and interact with automobiles in the future.

Data is at the core of developing the intelligence for the current and perceived future of ground vehicles. Data serves as the foundational resource for training and refining machine learning models that power various AI applications in vehicles. The vast amounts of data collected from sensors and other sources provide crucial insights into the complex dynamics of real-world mobility. This data allows AI models to learn and adapt, improving their ability to accurately perceive and respond to diverse and unpredictable conditions on the road. Moreover, data is instrumental in addressing safety concerns, enabling AI systems to understand rare events and edge cases that are critical for ensuring robust performance in varying situations. Consequently, data diversity is essential to the development of robust and accurate systems. The continuous collection and analysis of data contribute to the ongoing evolution of AI in ground vehicles, fostering innovation, enhancing reliability, and ultimately shaping the future of intelligent transportation systems.

SAE Executive Standards Committee Rules provide that: “This report is published by SAE to advance the state of technical and engineering sciences. The use of this report is entirely voluntary, and its applicability and suitability for any particular use, including any patent infringement arising therefrom, is the sole responsibility of the user.”

SAE reviews each technical report at least every five years at which time it may be revised, reaffirmed, stabilized, or cancelled. SAE invites your written comments and suggestions.

Copyright © 2024 SAE International

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, or used for text and data mining, AI training, or similar technologies, without the prior written permission of SAE.

TO PLACE A DOCUMENT ORDER: Tel: 877-606-7323 (inside USA and Canada)
Tel: +1 724-776-4970 (outside USA)
Fax: 724-776-0790
Email: CustomerService@sae.org
http://www.sae.org

SAE WEB ADDRESS:

For more information on this standard, visit
https://www.sae.org/standards/content/J3298_202407

The impact of data on ground vehicles is profound and multifaceted, influencing various aspects of vehicle operation, safety, and efficiency. All the different types of data could be used for two broad categories of ground vehicle use cases:

- a. To improve the safety, public acceptance, security, privacy, and performance of the ground vehicle itself
- b. To improve the traffic management, cooperative driving, navigation, and environmental impact reduction, etc., that involves the sharing of data and information between multiple traffic participants and infrastructure

Given the importance of data in the development of AI systems, it is important that stakeholders in the ground vehicle domain are knowledgeable of the current state of the industry as relates to data and AI systems. Stakeholders should be aware of some of the common types of datasets that currently exist, the types of licenses that may govern the use of certain datasets, methods and protocols for processing data, and challenges and issues associated with collecting, processing, and retaining data. This document provides information on each of the aforementioned topics.

- General Information Regarding AI and Data for Ground Vehicles

Data for AI development in ground vehicles is generated through a combination of onboard sensors, such as cameras, radar, lidar, and other advanced technologies. These sensors capture information about the vehicle's surroundings, including road conditions, traffic patterns, driver conditions and patterns, pedestrian movement, and potential obstacles. Additionally, vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication systems contribute to the generation of real-time data, enhancing the vehicle's situational awareness.

The collected data is then processed and used to train machine learning models. Machine learning models may be trained via supervised learning, which involves using labeled datasets, where the AI system learns from examples with known outcomes. Unsupervised learning may also be employed to identify patterns and relationships within the data without predefined labels. The continuous feedback loop of data collection, model training, and refinement is crucial for improving the accuracy and adaptability of AI systems.

Furthermore, data generated during actual driving scenarios helps AI models understand and respond to diverse and challenging conditions, contributing to the development of ADAS and automated driving capabilities. This data-driven approach enables vehicles to learn from real-world experiences, increasing safety measures and enhancing overall performance. As technology evolves, the collaborative sharing of anonymized and aggregated data among vehicles and stakeholders also plays a role in advancing AI development in the ground vehicle domain.

- Variances and Discrepancies in AI-Related Data

The ground vehicle domain presents a uniquely challenging landscape for AI, filled with variances and discrepancies in data. Environmental factors like everchanging weather, diverse road conditions, and varying lighting all significantly affect sensor inputs, causing inconsistency in data quality. Additionally, the quality and accuracy of data may be influenced by sensor calibration, maintenance issues, or inherent limitations in the sensing technologies. Furthermore, ground truths, often collected manually, can suffer from subjectivity and human error, introducing further discrepancies. These variations impact both training and performance, as AI models may struggle to generalize from limited data, produce erroneous results from biased datasets, or misinterpret nuanced data patterns. Addressing these challenges requires sophisticated data preprocessing techniques, robust sensor fusion strategies, validation, and careful analysis of ground truth data to mitigate bias and ensure that AI models in ground vehicles are capable of reliable performance across a broad spectrum of real-world conditions.

Many organizations employ standardized internal processes for data management when developing AI systems. However, many researchers, developers, and organizations in the ground vehicle domain are developing AI systems with little to no standard practice across the industry. While the current state of development may be beneficial to a single entity's progress, it could be detrimental the overall progress of AI systems in vehicles. A lack of standard practices, data formats, or recommended processes will make it difficult for vehicles developed by separate organizations to share information between vehicles or with infrastructure. Establishing industry-wide standards can ensure compatibility and interoperability among different AI systems. Developing data governance frameworks that outline the ethical and legal considerations could help address issues related to data privacy, consent, and the responsible use of AI technologies in ground vehicles. Establishing consistent, quantifiable metrics and pushing toward explainability in AI systems in this domain will require actions such as partnerships between the public and private sectors, regulatory oversight, open data initiatives, and continuous evaluation and updating.

TABLE OF CONTENTS

1.	SCOPE.....	4
2.	REFERENCES.....	4
2.1	Applicable Documents.....	4
2.1.1	SAE Publications.....	4
2.1.2	Other Publications.....	4
2.2	Related Publications.....	4
2.2.1	SAE Publications.....	4
2.2.2	ANSI Accredited Publications.....	5
2.2.3	Digital Governance Council Publications.....	5
2.2.4	DIN Publications.....	5
2.2.5	IEEE Publications.....	5
2.2.6	ISO Publications.....	6
2.2.7	NIST Publications.....	6
3.	TYPES OF STAKEHOLDERS.....	6
3.1	Original Equipment Manufacturers.....	6
3.2	Tier 1 Suppliers.....	6
3.3	Government Agencies.....	6
3.4	Legislators.....	7
3.5	Infrastructure Owners and Operators.....	7
3.6	General Public.....	7
3.7	Other AI Committees.....	7
3.8	Academia.....	7
4.	TYPES OF GROUND VEHICLE DATA.....	7
4.1	Data Collected by the Vehicle.....	8
4.1.1	Interior Data.....	8
4.1.2	Exterior Data.....	8
4.2	Vehicle-to-Everything (V2X) Data.....	9
5.	GROUND VEHICLE DATA RESOURCES.....	9
5.1	Real Datasets.....	10
5.2	Synthetic Data Sets.....	11
6.	PROCESSES AND PROTOCOLS.....	12
6.1	Data Collection Methods and Issues.....	12
6.2	Data Processing Methods and Issues.....	13
6.2.1	Preprocessing Methods.....	13
6.2.2	Postprocessing Methods.....	15
6.3	Unsupervised Learning.....	15
6.4	Data Management.....	16
6.4.1	Data Cards.....	16
6.4.2	Model Cards.....	16
6.5	Privacy and Security.....	17
7.	SUMMARY AND SUGGESTIONS.....	17
8.	NOTES.....	18
8.1	Revision Indicator.....	18
APPENDIX A	19

1. SCOPE

This SAE Technical Information Report provides preliminary information regarding the current state of data collection, data processing methods, and usage for developing AI and its enabled systems and applications in the ground vehicle domain. This information report is a survey of topics highlighting data's impact on AI solutions and methods that may be used to develop or improve data-related processes.

This report may offer insights that can drive innovation, improve safety, optimize performance, and develop regulatory compliance methods. Solution providers may find this information insightful and realize the potential for collaborative measures in development of their AI-enabled systems. Developers of standards and lawmakers may gain a better understanding of the current state of the industry and find opportunities to develop policies to guide the future of transportation, which in turn directly impacts the public. Other committees and academic affiliates may see links between data for AI in the ground vehicle domain and their domains of interest, which could spawn novel research, development, and standards.

NOTE: While this document is a survey of various topics that pertain to data in AI for ground vehicles, it largely applies to data that is geared toward vehicle navigation and monitoring surroundings that are external to the vehicle.

2. REFERENCES

Section 2 provides other documents that are related to data for AI and is a temporary placeholder for said documents for future discussions.

2.1 Applicable Documents

The following publications form a part of this specification to the extent specified herein. Unless otherwise indicated, the latest issue of SAE publications shall apply.

2.1.1 SAE Publications

Available from SAE International, 400 Commonwealth Drive, Warrendale, PA 15096-0001, Tel: 877-606-7323 (inside USA and Canada) or +1 724-776-4970 (outside USA), www.sae.org.

SAE J2735 V2X Communications Message Set Dictionary

2.1.2 Other Publications

Agarwal, S., Vora, A., Pandey, G., Williams, W., Kourous, H., and McBride, J. (2020). Ford multi-AV seasonal dataset. *The International Journal of Robotics Research*, 39(12), 1367-1376. <https://doi.org/10.1177/0278364920961451>.

Patil, A., Malla, S., Gang, H., and Chen, Y.-T. (2019, May 20-24). *The H3D dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes* [Conference paper]. 2019 IEEE International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada.

2.2 Related Publications

The following publications are provided for information purposes only and are not a required part of this SAE Technical Report.

2.2.1 SAE Publications

Available from SAE International, 400 Commonwealth Drive, Warrendale, PA 15096-0001, Tel: 877-606-7323 (inside USA and Canada) or +1 724-776-4970 (outside USA), www.sae.org.

SAE J3116 Active Safety Pedestrian Test Mannequin Recommendation

SAE J3157 Active Safety Bicyclist Test Targets Recommendation

SAE J3224 V2X Sensor-Sharing for Cooperative and Automated Driving

2.2.2 ANSI Accredited Publications

Copies of these documents are available online at <https://webstore.ansi.org/>.

ANSI/CTA-2090 The Use of Artificial Intelligence in Health Care: Trustworthiness

2.2.3 Digital Governance Council Publications

Available from Digital Governance Council, 1000 Innovation Drive, Suite 500, Ottawa, ON K2K 3E7, <https://dgc-cgn.org/>.

CAN/CIOSC 101:2019 Ethical design and use of automated decision systems

2.2.4 DIN Publications

Copies of these documents are available online at <https://www.din.de/en/>.

DIN SPEC 92001-1:2019 Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 1: Quality Meta Model

DIN SPEC 92001-2:2019 Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 2: Robustness

2.2.5 IEEE Publications

Available from IEEE Operations Center, 445 and 501 Hoes Lane, Piscataway, NJ 08854-4141, Tel: 732-981-0060, www.ieee.org.

IEEE P2975.1 Standard for Industrial Artificial Intelligence (AI) Data Attributes

IEEE 1232.1-1997 IEEE Standard for Artificial Intelligence Exchange and Service Tie to All Test Environments (AI-ESTATE): Data and Knowledge Specification

IEEE 2801-2022 IEEE Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence

IEEE 2802-2022 IEEE Standard for Performance and Safety Evaluation of Artificial Intelligence Based Medical Devices: Terminology

IEEE 2937-2022 IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems

IEEE 2941.1-2022 IEEE Standard for Operator Interfaces of Artificial Intelligence

IEEE 3129-2023 IEEE Standard for Robustness Testing and Evaluation of Artificial Intelligence (AI)-based Image Recognition Service

IEEE 3300-2022 IEEE Standard Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Technical Specification Multimodal Conversion Version 1.2

IEEE 3303-2023 IEEE Standard Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Technical Specification Compression and Understanding of Industrial Data 1.1

IEEE 3304-2023 IEEE Standard for Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Technical Specification Neural Network Watermarking (NNW) V1

IEEE 3806-2023 IEEE Standard for Blockchain-Based Hepatobiliary Disease Data Extraction and Exchange

IEEE 7010-2020 IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being

2.2.6 ISO Publications

Copies of these documents are available online at <https://webstore.ansi.org/>.

ISO/IEC CD TR 5469	Artificial intelligence – Functional safety and AI systems
ISO/IEC DIS 5259-1	Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 1: Overview, terminology, and examples
ISO/IEC DIS 5259-2	Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 2: Data quality measures
ISO/IEC DIS 5259-3	Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 3: Data quality management requirements and guidelines
ISO/IEC TR 24029	Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview
ISO/IEC TR 24028	Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence
ISO/IEC TS 25058:2024	Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Guidance for quality evaluation of artificial intelligence (AI) systems
ISO/IEC 22989:2022	Information technology – Artificial intelligence – Artificial intelligence concepts and terminology
ISO/IEC 23053	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
ISO/IEC 38507	Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations

2.2.7 NIST Publications

Available from NIST, 100 Bureau Drive, Stop 1070, Gaithersburg, MD 20899-1070, Tel: 301-975-6478, www.nist.gov.

NIST AI 100-1 Artificial Intelligence Risk Management Framework (AI RMF 1.0)

3. TYPES OF STAKEHOLDERS

3.1 Original Equipment Manufacturers

This document may be useful to original equipment manufacturers (OEMs) as they look to understand the current state of data usage and its impact on current and future production of their vehicles. The information provided may help OEMs realize approaches or limitations that could influence design, development, and production of their current and future systems.

3.2 Tier 1 Suppliers

This document may be useful to Tier 1 suppliers who develop hardware and software solutions for ground mobility applications. The information provided may help Tier 1 suppliers realize approaches or limitations that could influence design, development, and production of their current and future systems.

3.3 Government Agencies

This document may be useful for Government Agencies who look to develop policies and provide consistent measures for data usage in the applications of ground mobility (e.g., NIST).

3.4 Legislators

This document may be useful for legislators as it assists in formulating informed policies and regulations that promote the safe and efficient application of AI in ground mobility, while also addressing data privacy and security concerns. Such legislation could ultimately lead to safer, more efficient vehicles and a more robust ground vehicle industry.

3.5 Infrastructure Owners and Operators

This document provides insights into the data behind the development and use of AI in the automotive sector, which can have direct implications for infrastructure planning and management. By understanding the data requirements and the potential impact of AI on vehicle performance and traffic patterns, infrastructure owners and operators may be able to make more informed decisions about infrastructure development and maintenance. This could lead to improved traffic flow, reduced congestion, and ultimately, a more efficient and safer transportation system.

3.6 General Public

This document may be useful to consumers in the ground mobility domain given that data collection and usage has an impact on the mobility of the public. Furthermore, the evolution of methods, policies, and regulations also impact the mobility of the public.

3.7 Other AI Committees

Other AI Committees may find this document useful as there may be overlapping interests, approaches, and measures that could influence the production of similar documents in different domains as well as advance other documents in the ground mobility domain (e.g., Recommended Practices and Standards).

3.8 Academia

Researchers in academia may be interested to know the current state of the industry as it pertains to data for AI such that novel research and proposals may be discussed and generated.

4. TYPES OF GROUND VEHICLE DATA

Ground vehicle data plays an important role in the advancement of AI for ground vehicles, encompassing a wide array of information sources that influence vehicle operations and performance. This section delineates the various types of ground vehicle data and their significance in the context of AI-driven technologies. Ground vehicle data encompasses different types of information collected from vehicles that operate on roads or other surfaces. Some common types of ground vehicle data include:

- a. **Telematics Data:** This includes vehicle location, speed, acceleration, braking, fuel consumption, engine diagnostics, and other performance-related information transmitted wirelessly from the vehicle to a central system.
- b. **Sensor Data:** Vehicles are equipped with sensors such as cameras, lidar, radar, and ultrasonic sensors to detect the surrounding environment, including other vehicles, pedestrians, obstacles, and road conditions.
- c. **Vehicle-to-Everything (V2X) Data:** This data involves vehicle to vehicle communication (V2V), vehicles infrastructure communication (V2I), vehicle to pedestrian communication (V2P), and vehicle to network communication (V2N), enabling cooperative driving, collision avoidance, and traffic management.
- d. **Onboard Computer Data:** Modern vehicles have onboard computers that store data related to vehicle operation, maintenance history, diagnostic trouble codes (DTCs), and system alerts.
- e. **Navigation and Mapping Data:** Ground vehicles rely on navigation systems and maps to provide route guidance, traffic information, points of interest, and real-time updates on road conditions.

- f. Fleet Management Data: Fleet operators collect data on vehicle usage, driver behavior, fuel consumption, maintenance schedules, and vehicle health to optimize fleet operations and improve efficiency.
- g. Environmental Data: Ground vehicles may collect data on environmental factors such as air quality, temperature, humidity, and noise levels to monitor pollution levels and assess environmental impact.
- h. Automated Vehicle Data: Automated vehicles generate vast amounts of data related to perception, decision-making, control commands, and interactions with the environment, which is crucial for safe and efficient automated driving.

These are just a few examples of the diverse types of ground vehicle data available. With advancements in technology, there is a continuous expansion of the scope and capabilities of data collection and analysis in the automotive industry.

4.1 Data Collected by the Vehicle

Ground vehicles accumulate data from both internal and external sources, each contributing distinct insights into vehicle behavior and surrounding conditions. This subsection examines the difference between data collected within the vehicle's interior and that acquired from external sources.

4.1.1 Interior Data

Pertaining to vehicle-centric metrics such as battery health, engine performance, diagnostics, and onboard sensor readings, interior data is typically obtained from onboard sensors, control units, and diagnostic systems embedded within the vehicle's architecture. Examples of interior data sources include:

- Inertial Measurement Unit (IMU): Measures vehicle acceleration, deceleration, and orientation to assess dynamic motion characteristics
- Engine Control Unit (ECU): Monitors engine performance, fuel injection timing, and exhaust emissions
- Transmission Control Module (TCM): Manages gear shifting, clutch engagement, and transmission fluid temperature
- Vehicle Health Monitoring System: Tracks various components' health status, detects faults or malfunctions, and issues diagnostic codes for troubleshooting

4.1.2 Exterior Data

Encompassing information gathered from the vehicle's surroundings, including traffic conditions, weather patterns, and road infrastructure, this category of data is essential for understanding the vehicle's operating context and making informed decisions based on real-time situational awareness. Exterior data sources may include onboard sensors, cameras, lidar, radar, and communication modules that interact with the vehicle's surroundings. Examples of exterior data sources include:

- Environmental Sensors: Measure ambient temperature, humidity, air pressure, and visibility to assess weather conditions
- Traffic Cameras: Capture images or videos of surrounding traffic flow, road signs, and traffic signals to monitor congestion and detect anomalies
- GPS Receivers: Receive signals from global navigation satellite systems (GNSS) to determine the vehicle's precise location and navigate to desired destinations
- V2X Communication Modules: Exchange data with nearby vehicles, roadside infrastructure, and traffic management centers to share information about traffic incidents, road closures, and emergency alerts

4.2 Vehicle-to-Everything (V2X) Data

V2X communication enables data exchange between vehicles (V2V), infrastructure (V2I), pedestrians (V2P), and networks (V2N), fostering a connected ecosystem for cooperative driving. By facilitating real-time communication and information sharing, V2X enhances ground vehicle operations by providing vehicles with contextual awareness and collaborative functionalities. V2X communication systems utilize various wireless communication technologies, such as Dedicated Short-Range Communications (DSRC), Cellular Vehicle-to-Everything (C-V2X), and Wi-Fi, to enable data exchange between vehicles and their surroundings. This external data utilization allows vehicles to leverage real-time information from various sources, including traffic updates, road hazards, and pedestrian movements. By integrating this data, vehicles can optimize navigation, collision avoidance, and traffic management systems, thereby improving safety, efficiency, and the overall driving experience. Furthermore, V2X data enrichment enhances vehicle awareness and decision-making capabilities, enabling vehicles to anticipate hazards, plan routes, and coordinate actions with other vehicles well in advance. Ultimately, this collaborative approach leads to smoother traffic flow, reduced congestion, and safer driving environments for all road users. The main types of data exchanged in V2X communication include the following:

- **Sensor data:** This includes data from cameras, lidar, radar, and other sensors used by vehicles and infrastructure to perceive their surroundings.
- **Map data:** This includes information about roads, traffic signs, and other features of the environment.
- **Vehicle data:** This includes information about the vehicle's state, such as its speed, position, and heading.
- **Infrastructure data:** This includes information about traffic signals, road closures, and other relevant information from infrastructure elements.

In all the V2X use cases, accurate and timely data exchange, data integrity, privacy, and security play a key role in determining the successful data exchange between different traffic participants and the infrastructure in an effective manner. This is important to support and further enhance the effectiveness of cooperative transportation systems.

5. GROUND VEHICLE DATA RESOURCES

Ground vehicle data resources play a pivotal role in the contemporary landscape of transportation and automotive industries. These resources encompass a diverse array of information generated by ground vehicles, ranging from basic telemetry data to advanced sensor readings and communication signals. Ground vehicle data presents a vast and dynamic resource, fueling innovation in diverse fields like self-driving technology, traffic management, and predictive maintenance. As vehicles become increasingly connected and equipped with sophisticated sensors, the volume and complexity of data generated on roads and highways continue to grow. Leveraging this wealth of information opens opportunities for enhancements in the ground vehicle industry, but it also presents challenges.

One challenge is that of data collection. Researchers and developers in the domain of AI in ground vehicles require vast amounts of the appropriate data to strategically train their system for a particular application. Additionally, training neural networks, specifically, requires large amounts of carefully annotated datasets. How can data scientists obtain the data they require, at the desired or acceptable size and rate, without sacrificing accuracy, balance, or quality? This section highlights two types of data resources that are often used in the development AI systems in ground vehicles.

5.1 Real Datasets

Real-world or real data represents authentic information collected in real-life use cases. Many real data sets are created by attaching sensors to vehicles and physically driving said vehicles in various environments to gather the data that is received via the attached sensors. For example, the Honda 3D Dataset (H3D) is a large scale full-surround 3D multi-object detection and tracking dataset. It is gathered from the HDD (Honda Research Institute Driving Dataset), a large-scale naturalistic driving dataset collected in San Francisco Bay Area,¹ and consists of lidar point clouds and object annotations from crowded and interactive traffic scenes.² Another example is Ford's Multi-Seasonal Automated Vehicle Dataset that is a multi-agent seasonal dataset collected by a fleet of Ford vehicles at different days and times during 2017 to 2018. The vehicles were manually driven on a route in Michigan that included a mix of seasonal variations in weather, lighting, construction, and traffic conditions experienced in dynamic urban environments.³

While some entities manually operate vehicles for data collection, others use more automated methods. For instance, the Dense Depth for Autonomous Driving (DDAD) dataset is an autonomous driving benchmark from the Toyota Research Institute (TRI) for long range and dense depth estimation in various urban conditions. It contains monocular videos and ground-truth depth generated from high-density lidars mounted on a fleet of automated cars operating in a cross-continental setting.⁴ Additionally, the Waymo Open Dataset is a diverse multimodal autonomous driving dataset that comprises of images recorded by multiple cameras and sensor readings from multiple lidars mounted on a fleet of automated vehicles. Like other datasets, it comprises of recordings across several conditions in multiple cities. The dataset also contains numerous manually annotated 3D bounding boxes for lidar data and 2D bounding boxes for camera images.⁵ To extend their dataset, researchers may perform data augmentation. Data augmentation is a modification of existing real-world data to extend the dataset and increase robustness in the trained neural network. For example, developers may perform augmentation of images by rotating or brightening an existing image to create a new one.

Entities may employ various techniques for collecting sensor data via vehicular travel. Often, it is prudent to consider a method that allows simultaneous capture of multiple sensors such that the data from each sensor is properly synchronized during the collection process. One approach is to use the robotics middleware suite called Robot Operating System (ROS). ROS is free and open-source software that defines the components, interfaces, and tools for building robotic systems. By enabling connections between sensors, actuators, and control systems via tools called topics and messages, ROS helps developers build robotic systems that can interact with the world. Developers can record messages using ROS bag files or logs for testing, training, or quality assurance, and those messages can contain pertinent information from hardware interfaces such as cameras, lidars, and motor controllers. Hence, ground vehicle researchers may use ROS, specifically ROS bag files, to collect data from sensors attached to vehicles.

Beyond data explicitly collected for automated driving, other data not originally intended for AI in ground vehicles may be used to train AI models. The second Strategic Highway Research Program (SHRP2) Naturalistic Driving Study, published via Virginia Tech Transportation Institute (VTTI), Transportation Research Board (TRB), and SHRP2, focuses on driver performance and behavior in traffic safety. As noted by the researchers, "this involves understanding how the driver interacts with and adapts to the vehicle, the traffic environment, roadway characteristics, traffic control devices, and other environmental features."⁶ Other data-rich projects may focus more heavily on improving the safety of travelers and pedestrians through connected vehicle (CV) efforts such as V2X technologies.⁷ The CV Pilot Deployment Program partnered with the Intelligent Transportation System (ITS) Data Program to make sanitized and anonymized data from the sites available on a public-facing portal, ITS DataHub.⁸ During this pilot program, multiple Departments of Transportation (DOT) shared event logs of relevant Basic Safety Messages (BSM), Traveler Information Messages (TIM), Map Data Message (MAP), and Signal Phase and Timing (SPaT) messages, which follow the SAE J2735 standard.⁹ Such information could be used to train an AI model to improve the drive or ride experience both internally and externally to the vehicle.

¹ <https://usa.honda-ri.com/h3d>

² Refer to Patil et al. (2019).

³ Refer to Agarwal et al. (2020).

⁴ <https://arxiv.org/abs/1905.02693>

⁵ <https://arxiv.org/pdf/1912.04838.pdf>

⁶ https://insight.shrp2nds.us/documents/shrp2_background.pdf

⁷ https://www.its.dot.gov/pilots/pilots_nycdot.htm

⁸ "Connected Vehicle Pilot (CVP) Open Data," ITS DataHub, [Online]. Available: <https://datahub.transportation.gov/stories/s/Connected-Vehicle-Pilot-CVP-Open-Data/hr8h-ufhg/>.

⁹ <https://rosap.ntl.bts.gov/view/dot/68128>

Regardless of the originally intended purpose of the collected data, those desiring to use publicly available data must consider the potential issues that surround all data. Issues such as biases in the data collection process and privacy or regulatory compliance concerns may present roadblocks and may have ramifications when using data to train AI models. Biases in publicly available datasets may go unnoticed but can cause significant issues within deployed systems. Regarding privacy or regulatory compliance, many publicly available datasets carry varying levels of Creative Commons (CC) license languages that limit the use of the data. A CC license is one of several public copyright licenses that enable the free distribution of an otherwise copyrighted “work.” A CC license is used when an author wants to give other people the right to share, use, and build upon a work that the author has created.¹⁰ The four types of CC licenses are:

- Attribution (BY): The user of the work must attribute the author of the work.
- Share Alike (SA): Any adaptations of the work must be licensed under the same or compatible license.
- Non-Commercial (NC): The work may only be used for non-commercial purposes.
- No Derivatives (ND): The user cannot share an adaption of the work, but the user can use and share the work in its original form.

A list of several real datasets and their associated CC license languages can be found in [Appendix A](#).

5.2 Synthetic Data Sets

Synthetic data is data that has been artificially created by computer algorithms imitating real data. Generally, generating synthetic data samples is more cost-efficient and less time-consuming than collecting real data from natural events. Entities typically use synthetic data as a method of increasing the number of the data points to train the models in their system. For example, the Carnegie Mellon University Robotics Institute developed the All-In-One Drive (AIODrive), which is a synthetic dataset that provides comprehensive sensors, annotations, and environmental variations geared toward aiding automated vehicle development.¹¹ The Virtual Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) Dataset, created by Naver Labs, is a synthetic dataset designed to learn and evaluate for video understanding tasks.¹² Other entities, such as BMW, are using synthetic data to improve the efficiency of their factories when building their vehicles.¹³ A list of several synthetic datasets and their associated CC license languages can be found in [Appendix A](#).

Developers of synthetic datasets generally use simulators to create said datasets. CAR Learning to Act (CARLA), for example, is an open simulator for urban driving, developed as an open-source layer over Unreal Engine.^{14,15} Microsoft Research created the open-source platform Air Sim as a simulator for software-in-the-loop development for drones and ground vehicles.¹⁶ Microsoft has now launched Project Air Sim, which focuses on accelerating efforts for autonomous flight.¹⁷ NVIDIA’s Omniverse Replicator is yet another platform that allows developers to build synthetic datasets for training AI models.¹⁸

Synthetic datasets and methods used to create them have the advantages of enabling large volumes of training data in a broad range of applicable fields. However, synthetic data also suffers from issues that are potentially harmful to the training and deployment processes for AI models. While users of publicly available synthetic datasets must also understand the rules and regulations regarding the use of said data (similar to those in real datasets), synthetic data may also suffer from biased or deceptive results, which can lead to misleading, limited, or discriminatory output due to a lack of variability or diversification. Additionally, synthetically generated data may suffer from unnatural or unrealistic physical output. Consequently, AI models may learn characteristics from the data that do not naturally occur in real world applications.

¹⁰ https://en.wikipedia.org/wiki/Creative_Commons_license

¹¹ <https://openreview.net/pdf?id=yI9aThYT9W>

¹² <https://doi.org/10.48550/arXiv.1605.06457>

¹³ <https://www.press.bmwgroup.com/global/article/detail/T0375993EN/bmw-group-publishes-sordi-the-largest-open-source-dataset-by-far-for-super-efficient-ai-applications-in-production?language=en>

¹⁴ <https://paperswithcode.com/dataset/carla>

¹⁵ <https://carla.org/>

¹⁶ <https://microsoft.github.io/AirSim/>

¹⁷ <https://news.microsoft.com/source/features/innovation/microsoft-launches-project-air-sim-an-end-to-end-platform-to-accelerate-autonomous-flight/>

¹⁸ <https://developer.nvidia.com/omniverse/replicator>

6. PROCESSES AND PROTOCOLS

To generate data sets, whether real or synthetic, entities generally follow certain processes and protocols that aid in their development process for their specific purpose. Such processes and protocols are typically designed to aid with the entity's consistency in capturing, storing, and using the data for training AI models with a core focus to produce a well-orchestrated workflow. It is worth discussing some of the standard processes, while also noting current challenges. The subsections below outline methods and issues within certain standard processes and protocols for data in AI for ground vehicles.

6.1 Data Collection Methods and Issues

Data collection is essentially the process of gathering information that is typically used to investigate, reason, and act upon for a given purpose. In the ground vehicle domain, a company may collect data to better understand the use cases of its vehicles. For AI-specific data, a company may collect data to train an AI model that provides statistical information and suggestions to drivers that may help with their commute. Data collection methods can be both led by internal processes or outsourced externally. Regardless, both processes require insurance of data quality, integrity, and consistency. Some methods and corresponding issues of collecting AI training data include:

a. Web Scraping

1. Method: The process (generally automated) of extracting data from online sources (e.g., websites).
2. Issues:
 - i. Post analysis is required after the data has been gathered to ensure accuracy and continuity.
 - ii. Cost of maintenance can be high due to frequent changes of website.
 - iii. Collected data can be limited due to anti-scraping tools.

b. Data Generation

1. Method: Creating data (often synthetically [see [5.2](#)]) that will be used to train AI models.
2. Issues:
 - i. Unnatural or unrealistic physical output.
 - ii. Hardware and software costs can be high.
 - iii. May lack social acceptance in certain domains.

c. Crowdsourcing

1. Method: The process of having the public create data generally via an entity-provided instructions and sharing platform.
2. Issues:
 - i. Data quality is challenging to track.
 - ii. Finding contributors qualified in the desired domain can be challenging.
 - iii. Ensuring proper adherence to the instructions can be challenging.

e. In-House (or Private) Data Collection (see [5.1](#))

1. Method: The process of collecting data via internal operations, often due to challenges such as sensitivity or domain-specificity.
2. Issues:
 - i. Hiring, recruiting, and/or collecting data internally is expensive and time-consuming.
 - ii. It can be difficult to gather information in a domain that is not well-researched for AI development.
 - iii. Instrumentation, time synchronization, and onboard storage are also practical problems.

6.2 Data Processing Methods and Issues

Raw data is generally not useful to any entity. After collecting raw data, organizations generally translate the data into useful information. Developers may perform actions such as filtering, sorting, processing, and/or storing before establishing the base format for the useable data. Developers may use software like ROS (see [5.1](#)) to process raw data before sending data to an AI algorithm for analysis and comprehension. After analyzing, the data may be processed again for further readability and use in defined scenarios. Hence, data processing methods include Data Preprocessing and Postprocessing for improved performance and attaining the desired outputs.



6.2.1 Preprocessing Methods

Data preprocessing is an important step in the data analysis process that involves techniques to change raw data into useable data before sending the data to an AI model for analysis. Often, raw data may contain inconsistencies, noise, or missing information that can corrupt an algorithm's performance. For example, sensor data such as radar data suffers from issues such as multipathing. Hence, preprocessing methods are deployed to aid in filtering data that is outside of the normal range of values.

Various methods such as handling missing data, normalization, data augmentation, encoding, scaling, noise reduction, transformation, synthetic data generation, text processing, dimensionality reduction, balancing data, and time series data handling, etc., aim to clean or sanitize, organize, and structure data, ensuring it is well-prepared and suitable for AI algorithms, enabling accurate and efficient model performance.

Some of the issues with data preprocessing include the following:

- a. Quality of Data – Accuracy and Completeness
- b. Volume of Data – Scalability and Storage
- c. Velocity of Data – Real-Time Processing and Streaming Data
- d. Variety of Data – Diverse Formats and Unstructured Data
- e. Veracity of Data – Data Quality Assurance and Data Bias
- f. Security and Privacy – Data Protection and Cybersecurity

6.2.1.1 Data Annotation Methods and Issues

Data annotation is an important part of training AI models in the ground vehicle domain. Properly labeled data sets are vital for learning algorithms to identify patterns and make accurate predictions. Below is an overview of data annotation methods:

- a. Bounding Box Annotation is a method that involves drawing rectangles around objects of interest in an image to define their location. The annotations may be used to identify and localize vehicles, pedestrians, and other objects on the road.
- b. Key Point Annotation is a method that involves marking specific points on an object of interest that will help in understanding the object's structure. The annotations may be used to identify key points on a vehicle, such as license plates, for vehicle identification and tracking.
- c. Semantic Segmentation is a method that involves labeling each pixel in an image with a specific class, allowing the model to understand the context of each pixel. The annotations may be used to identify road lanes, understand the road environment, and differentiate between objects.
- d. Instance Segmentation, similar to semantic segmentation, is a method that distinguishes between individual instances of the same class. The annotations may be used to identify and track multiple instances of the same type of object, such as multiple vehicles.
- e. Temporal Annotation is a method of annotating data with temporal information to train models in understanding movement and predicting future actions. The annotations may be used to predict the trajectory of vehicles or vulnerable road users.
- f. Three-Dimensional Annotations allow for information such as depth, distance, and volume to be factored into the annotation process. Examples of 3D annotations are cuboids and voxels (3D pixels). The annotations may be used to identify and localize vehicles, pedestrians, and other objects on the road and are often captured via depth-measuring sensors like lidar, radar, or stereo cameras.

Below are some associated issues with data annotation processes:

- a. Subjectivity and Variability: Inconsistent labeling due to various interpretations of the data between different annotators.
- b. Lack of Standardization and Protocols: Non-standardized annotation practices or guidelines can result in varied annotations.
- c. Scale and Occlusion: Annotating the data may be challenging due to object variances such as size or occlusion.
- d. Imbalanced Datasets: Poor distribution of classes or lack of data diversity may lead to biased annotation and biased models.
- e. Time-Intensive and Costly: Manual (human-based) annotation can be time-consuming and expensive and can introduce errors due to fatigue in the annotators. Semi-automated annotation can alleviate the time associated with the annotation process, but errors may persist as the automated and non-automated portions may still yield poorly labeled data.
- f. Privacy Concerns: Certain information such as license plates or faces may raise privacy concerns during the annotation process and may require careful handling of data.
- g. Continuous Updates: Managing and updating annotations as technology evolves may be a logistical challenge.

Addressing these challenges involves strategic planning that includes quality control measures for the development of advanced AI algorithms. As the field advances, ongoing efforts are required to improve the efficiency and accuracy of data annotation processes.

6.2.2 Postprocessing Methods

Often, AI algorithms produce outputs that are less than ideal for the intended end use. For example, specifically in image processing, some AI models do not produce full images as outputs, rather they produce quantifications within the images such as object detections in image coordinates. In some cases, image coordinates will suffice as an output. However, it is often beneficial to produce an image with the object coordinates identified for visual inspection. Therefore, postprocessing techniques are deployed for human interpretability.

Various postprocessing methods include thresholding, filtering, normalization, ensembling, calibration, error analysis, visualization, interpretability techniques, and feedback loop implementation. They aim to refine and interpret AI model outputs, ensuring reliability, interpretability, and improved performance metrics tracking over certain hours of operation or miles of operation.

Issues with data postprocessing include the following:

- a. Security and Privacy – Data Protection and Cybersecurity
- b. Interpretability and Explainability – Black Box Models and Model Transparency

Addressing these challenges requires a comprehensive approach, combining technological solutions, ethical considerations, and adherence to regulatory frameworks.

6.3 Unsupervised Learning

A current trend, at the time of this report, is unsupervised learning. Unsupervised learning is an AI process where the algorithm is given input data without explicit instructions on what to do with it. The system tries to learn the patterns and structure from the data without the specific guidance of human-labeled examples. Some examples of unsupervised learning techniques include clustering, anomaly detection, self-organizing maps, and general adversarial networks (GANs). Arguably, the most prominent topic, currently, is self-supervised learning, which is a specific approach within unsupervised learning and is often used to develop generative AI systems. In self-supervised learning, an algorithm generates its own labels or supervisory signals from the input data. Generally, developers of self-supervised learning algorithms generate a pretext task (encouraging the model to learn useful representations or features from the input data) to train a model to solve this pretext task, which can then be used as the foundation for further learning via supervised, semi-supervised, or even more unsupervised methods.

Unsupervised and self-supervised learning span the entirety of the data processing pipeline. An example of self-supervised learning within the ground vehicle domain could be to train a vehicle model to anticipate its own future states and actions, given past observations. This process would require researchers and developers to first collect sequences of observation data from the vehicle's sensors. Pertinent information could be images, point clouds, and vehicle state. Researchers and developers would then conduct a preprocessing step to align, synchronize, and augment the data for variability and robustness. The pretraining process would then include the construction of a neural network to process the data, design of the pretext task to accept past observations as inputs and predict a future vehicular state, and training the model to minimize prediction errors via a mathematical model. Researchers and developers may choose to fine-tune the model on labeled data from specific tasks, adjusting its parameters to improve performance on the target tasks such as obstacle detection.

This process carries significant advantages such as robustness to feature learning and domain adaptation. In general, it also reduces the manual annotation process. Still, the process maintains its own set of challenges. The process heavily relies on the choice and design of the pretext task. Models will extract their own pertinent features from the data provided by means of the pretext task, which challenges developers to generate meaningful pretext tasks and data. Consequently, just like previously mentioned algorithms, the data could still lack diversity, which could produce models that are overfit to the pretext task and yield non-generalizable and biased models. Further, the current development of these systems requires large amounts of unlabeled data and substantial computational resources for training complex neural network architectures.

6.4 Data Management

Data management is process of collecting, organizing, securing, and storing data. Data is arguably one of the most important resources of an organization, and it is critical for entities to ensure they are employing methods to properly maintain their data. The goal of data management is to optimize the use of the data for its intended purpose while adhering to policies and regulations that govern the data. While organizations may have their own internal policies for data management, there currently does not exist standard policies or regulations across the ground vehicle domain. Such standards may make it more seamless for organizations to share information, which could be beneficial for the industry as a whole. Some organizations have begun to publish methods for managing data and machine learning models that may be adopted as recommended practices in the future.

6.4.1 Data Cards

Researchers at Google created and published an approach to provide information about a specific dataset that may be useful to potential users of said dataset. As defined by the authors, data cards are structured summaries of essential facts about various aspects of ML datasets needed by stakeholders across a project's lifecycle for responsible AI development.¹⁹ As the researchers note, as AI model development and complexity increases, a clear and thorough understanding of a dataset's origins, development, intent, ethical considerations, and evolution becomes a necessary step for the responsible and informed deployment of models.²⁰ In the publication, researchers present information regarding their development methodology, introduce a transparency artifact for production and research environments, propose three frameworks for producing data cards, and present lessons learned from deploying over 20 data cards. A single data card may include information about the author of the dataset, motivation for creating the dataset, the intended use of the data, and even partial examples of the data itself, which may highlight formatting information. A data card may include other pertinent information such as the annotation process, licensing, and versioning.

As an example, someone may wish to create a data card for a dataset that pertains to object detection at an intersection. The data card would begin with a brief description of the purpose of the dataset, potentially including relevant information like the number of images and sensor type as well as how the data was annotated. The data card would then provide information about the authors of the dataset, such as the publisher's name, funding source, and industry type. Next, the data card would provide more detailed information regarding the motivation for creating the dataset, listing pertinent information such as key applications and the problem space. The motivation would be followed by information about the use of the dataset, focusing on details of the applicable space as well as non-applicable spaces, and the method for using the data (i.e., object detection). The data card would go on to provide a numerical snapshot of the dataset, listing information such as the total number of images, the total number of labels, and a description of the annotations. The data card may even provide an example of a datapoint (e.g., list the specific image coordinates of a label and the label type). The data card would then provide information about the method(s) used to collect the data, such as data being collected during daylight hours via local cameras mounted at traffic intersections. Finally, the data card would provide detailed information about the labeling process, focusing on aspects such as manual or automated labeling, label attributes (e.g., person or vehicle and vehicle type), and a further numerical breakdown of each label category (e.g., N number of people, M number of vehicles, X number of cars, Y number of trucks, etc.).

6.4.2 Model Cards

In a similar fashion to the data cards, researchers at NVIDIA created and published a potential approach to document ML models, termed model cards. The intent of model cards is for developers to disclose the intended use cases for their ML models, highlighting relevant information such as performance metrics, and minimize the use of models in non-applicable contexts. As defined by the authors, model cards are short documents accompanying trained ML models that provide benchmarked evaluation in a variety of conditions that are relevant to the intended application domains, and model cards disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information.²¹ In the publication, researchers present examples model cards for two models: one image-based classification and one text-based scoring system.

¹⁹ <https://arxiv.org/pdf/2204.01075.pdf>

²⁰ <https://arxiv.org/pdf/2204.01075.pdf>

²¹ <https://arxiv.org/pdf/1810.03993.pdf>

Using the same example presented in [6.4.1](#), a model card for object detection at an intersection would begin by providing details of the model itself, listing information about the developers, specific type of model (e.g., CNN, RNN, etc.), and a high-level description of the training process. Next, the model card would provide information about the intended use of the model (e.g., intended to be use for crash mitigation at intersections) as well as uses not suitable for the model. The model card would then list some known factors that pertain to the trained model, which may include information such as data limitations in the training process or even physical limitations with the sensors themselves. The model card would also provide metrics of the model, such as false positive and false negative rates. The model card would also provide information about the training data and evaluation data used for the model, which, in this case, could be tied directly to the data card presented previously. Finally, the model card would list any ethical considerations, such as data being captured in a public space, and any caveats or recommendations, which could relate to the ethical considerations.

6.5 Privacy and Security

Privacy and security concerns are paramount in the management of data in AI for ground vehicles. The integration of AI in vehicles, encompassing features like ADAS and automated driving, necessitates the collection and analysis of vast amounts of potentially sensitive data. This data often includes real-time location information, sensor readings, and potentially identifiable details about individuals. Ensuring the privacy of individuals is a critical challenge, demanding robust measures such as anonymization and encryption to safeguard against unauthorized access or malicious use. Unauthorized access to this data could lead to invasive surveillance or compromise personal information. Moreover, the connectivity of AI-enabled vehicles introduces the risk of cyber threats and malicious interference that could result in devastating consequences, from accidents to coordinated attacks. Consequently, the industry will need to emphasize stringent cybersecurity protocols to mitigate hacking attempts and protect both user privacy and the overall safety of the transportation ecosystem. Balancing between innovative solutions and safeguarding personal data remains a continuous challenge in the evolution of AI for ground vehicles.

7. SUMMARY AND SUGGESTIONS

In the rapidly evolving landscape of automotive technology, the integration of AI has emerged as a transformative force, particularly in ground vehicles. Consequently, the data associated with AI systems is vital in the training and development of the groundbreaking advancements in safety, efficiency, and overall driving experience. This information report delves into the pivotal role that data plays in powering AI systems within ground vehicles, exploring the diverse applications and challenges that shape the connection of data and AI.

As discussed in this information report, several entities are deploying various methods for collecting and managing their data to maintain a process for training their models for their particular use cases. Each approach has merit, and no one approach currently fits the masses. However, without consistent approaches or industry standards associated with the development of modern AI systems for ground vehicles, challenges may arise. Nonexistent or even inconsistent standards may lead to varying levels of safety in AI-enabled vehicles. Without a unified approach, there is a risk of inadequate safety measures and protocols, potentially resulting in malfunctions or accidents. The absence of consistent standards may complicate the regulatory environment for AI in ground vehicles. Governments and regulatory bodies may find it challenging to create and enforce rules given the technology lacks standardized practices.

Furthermore, the future of the ground vehicle industry will likely include a vast network of V2X communication. As noted previously in this document, there currently is no industry standard for managing and sharing data or models. A lack of standards can lead to interoperability issues between different AI systems. Vehicles equipped with diverse AI technologies may struggle to communicate effectively, hindering collaboration and coordination in mixed traffic scenarios. Additionally, inconsistent approaches may result in varied security protocols. This could make AI systems more susceptible to cyber threats, increasing the safety concerns of the public. Data cards and model cards may be foundational frameworks towards standard practices for the ground vehicle industry to consider.

From enhancing automated driving capabilities to optimizing vehicle performance, the relationship between data and AI is reshaping the future of transportation, promising a new era of intelligent and connected vehicles. Inconsistency in AI systems could erode public trust in the development and deployment of vehicle technology. Without standardized safety measures and transparent practices, people may be skeptical about the reliability and safety of AI-equipped ground vehicles. To address these challenges, it is critical for the industry to work collaboratively to create and administer consistent approaches and standards in the development and deployment of AI systems for ground vehicles. This includes efforts from technology developers, regulators, and other stakeholders to ensure the safe, ethical, and effective integration of AI into transportation systems.

Given the current state of the ground vehicle domain and the likely continued growth of incorporating AI algorithms in the industry, this taskforce suggests the development of a recommended practice for data used to implement AI algorithms. Such recommended practice could subsequently lead to an industry standard that may address each of the challenges discussed in this document.

8. NOTES

8.1 Revision Indicator

A change bar (I) located in the left margin is for the convenience of the user in locating areas where technical revisions, not editorial changes, have been made to the previous issue of this document. An (R) symbol to the left of the document title indicates a complete revision of the document, including technical revisions. Change bars and (R) are not used in original publications, nor in documents that contain editorial changes only.

PREPARED BY SAE ARTIFICIAL INTELLIGENCE

Copyright © SAE International

APPENDIX A

A.1 EXAMPLES OF REAL DATA SETS

Real Data Sets														
S.no.	Dataset Name	Creator	Year	License	Classes	Scenes	Camera Type(s)	RGB Images	Lidar Type	Lidar PCs	Radar Type	Radar PCs	Data Retrieval Date	Reference
1	Ford Multi-AV Seasonal Dataset	Ford	2020	CC BY-NC-SA 4.0	-	32* (8 x 4)	6 - 1.3MP 1 - 5.0MP	-	4 - 32 Beam	-	no Radar	-	9/7/2023	https://doi.org/10.48550/arXiv.2003.07969
2	ONCE Dataset (One million sCenEs)	Huawei Corporation	2021	CC BY-NC-SA 4.0	5	1M	7 - 2.0MP	7M	1 - 40 Beam	1M	no Radar	-	9/7/2023	https://doi.org/10.48550/arXiv.2106.11037
3	Dense Depth for Autonomous Driving (DDAD)	Toyota Research Institute	2020	CC BY-NC-SA 4.0	-	200	6 - 2.4MP	99.6K	1 - 64 Beam	-	no Radar	-	9/7/2023	https://doi.org/10.48550/arXiv.1905.02693
4	KITTI-360 Dataset	Karlsruhe Institute of Technology and Toyota Technological Institute	2022	CC BY-NC-SA 3.0 DEED	19	4x83,000	Fisheye	320k	1 - 64 Beam	100k	no Radar	-	2/13/2024	https://arxiv.org/pdf/2109.13410.pdf
5	PandaSet	Scale	2020	CC BY 4.0	28* 37*	100+	6 -	48k+	1 - 64 Beam 1 - Solid State	16k+	no Radar	-	9/7/2023	https://doi.org/10.48550/arXiv.2112.12610
6	A2D2: Audi Autonomous Driving Dataset	Audi	2020	CC BY-ND 4.0	38	-	6 - 2.3MP	40k+*	5 - 16 Beam	40k+*	no Radar	-	9/7/2023	https://doi.org/10.48550/arXiv.2004.06320
7	Canadian Adverse Driving Conditions (CADC)	The University of Waterloo	2020	CC BY-NC 4.0	10	75	8 - 1.3MP	56K	1 - 32 Beam	7K	no Radar	-	9/7/2023	https://doi.org/10.48550/arXiv.2001.10117
8	H3D - Honda 3D Dataset	Honda	2019	CC BY-NC-SA 4.0	8	160	NA	NA	1 - 64 Beam	1M*	no Radar	-	9/7/2023	https://doi.org/10.48550/arXiv.1903.01568
9	nuImages	nuScenes	2020	CC BY-NC-SA 4.0	23	1K	6 - 1.9MP	93k* 1.2M	1 - 32 Beam	390K	5 - 77GHz	1.4M	9/7/2023	https://doi.org/10.48550/arXiv.1903.11027
10	Waymo Open	Waymo	2023	open dataset	-	1150(12 million 3D labels and 1.2 million 2D labels), 200k camera frames	2.5 MP(S front and side facing cameras)	23k* 9.0M	5 high resolution Lidars	12M, 64 channels	no Radar	-	10/28/2023	https://github.com/waymo-research/waymo-open-dataset
11	Berkeley Deep Drive Dataset	Berkeley	2020	-	10 [19 classes evaluated for semantic segmentation]	100k videos(40 sec scenes each),	0.9MP	318k+	No Lidar used	-	no Radar	-	10/28/2023	https://doi.org/10.48550/arXiv.1805.04687
12	Cityscapes Dataset	Cityscapes	2020	-	30	-	Stereo RGB images	ted images with fine annotations), 20k(annotated images with coarse annotations)	No Lidar used	-	no Radar	-	10/28/2023	https://doi.org/10.48550/arXiv.2006.07864

A.2 EXAMPLES OF SYNTHETIC DATA SETS

Synthetic Data Sets															
S.no.	Dataset Name	Creator	Year	License	Classes	Scenes	Ann. Frames	RGB Camera Type	RGB Images	Lidar Type	Lidar PCs	Radar Type	Radar PCs	Data Retrieval Date	Reference
1	AIODrive	CMU Robotics Institute	2020	CC BY-SA 4.0	23	-	100K	5 - 1.4MP	-	3 - 64/800/1200 1 - SPAD	100k/600k/1M per frame	4 - 10hz	1.5M	8/31/2023	https://www.xinshuoweng.com/papers/AIODrive/supp.pdf
2	SHIFT	ETH Zurich, MPI Informatics, Google, Technical University of Munich	2022	Open Dataset	13	-	2.5M	1280 × 800 pixel	2.5M	128 channel	10hz	no Radar	-	10/28/2023	https://arxiv.org/pdf/2206.08367.pdf
3	PerSIL	University of Waterloo and University of Toronto, Canada	2019	Open Dataset	20+	25k video frames	50K	camera[1920x1080 pixel] (within game simulator Grand Theft Auto V)	200k	Synthetic Lidar to mimic Velodyne HDL-64E	-	no Radar	-	10/28/2023	https://arxiv.org/pdf/1905.00160.pdf
4	OPV2V	UCLA Mobility Lab and Cleveland State University Vision and AI Lab	2022	Open Dataset	15+	11,464 frames	11K , 232K annotated 3D box diagrams	4 RGB cameras	-	64 channel	1.3M points per sec	no Radar	-	10/28/2023	https://arxiv.org/pdf/2109.07644.pdf
5	SYNTHIA	UAB Barcelona, University of Vienna	2016	Open Dataset	30+(11 classes ex. Shown)	50K - 200K frames	-	2.1 MP, 8 RGB Cameras used(FOV 100 deg each)	213k	No Lidar	-	no Radar	-	10/28/2023	https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Ross-The_SYNTHIA_Dataset_CVPR_2016_paper.pdf
6	Synscapes	7D labs, Linköping University, Sweden	2018	Open Dataset	18	-	-	0.9 & 2.1MP (1440x720,2048x1024)	25k	No Lidar	-	no Radar	-	10/28/2023	https://arxiv.org/pdf/1810.08705.pdf
7	VIPER	KAIST, Adobe Research	2020	Open Dataset	31	184k	2300 Frames available and 79 unique tracks were	2.1 MP	-	No Lidar	-	no Radar	-	10/28/2023	https://arxiv.org/pdf/2006.11339.pdf
8	V2X-Sim	USC, LA, Shanghai Jiao Tong University	2022	Open Dataset	7+	100	10K	1600x900	-	Lidar with 360 deg coverage	32 channels, 70 m max range, 250k points per sec	no Radar	-	10/28/2023	https://arxiv.org/pdf/2202.08449.pdf
9	CARLA	University of Patras, Panasonic Automotive, Germany	2021	Open Dataset	23, [5 main classes and subclassifications within the classes]	-	-	1280x960	30 fps	2 Lidars (velodyne)	1x16,1x64 channels	no Radar	-	10/28/2023	https://openaccess.thecvf.com/content/CVPR2022W/WAD/papers/Kloukiniotis_CarlaScenes_A_Synthetic_Dataset_for_Odometry_in_Autonomous_Driving_CVPRW_2022_paper.pdf
10	VEIS	ANU, CSIRO, ACRV, CVLab, EPFL, NVIDIA	2018	Open Dataset	19	61,305 frames	31125 frames for single class	Virtual camera	61k	No Lidar	-	No Radar	-	10/28/2023	https://openaccess.thecvf.com/content/ECCV_2018/papers/Fatemeh_Sadat_Saleh_Effective_Use_of_ECCV_2018_paper.pdf