Responsible Generative and Agentic Artificial Intelligence Frameworks for Autonomous Electric Vehicle Adoption

Abhinav Tiwari

A dissertation submitted to the Faculty of Graduate Studies

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Graduate Program in

Electrical Engineering and Computer Science

York University, Toronto, Ontario

January, 2026

Abstract

Electrified and autonomous mobility are jointly reshaping transportation and the supporting energy and data infrastructures. Electric Vehicles (EVs) introduce new load patterns, infrastructure constraints, and opportunities for mobility assets to serve as flexible distributed energy resources (DER). Autonomous Vehicles (AVs) intensify reliance on artificial intelligence (AI) for safety-critical perception and decision-making, amplifying the significance of AI risks around system failures, bias, security vulnerabilities, and opaque behaviour. The increased digitization of mobility further elevates the need for strong data governance and privacy-preserving data sharing, as AI, including generative and agentic models, amplifies both automation's benefits and associated risks.

This thesis advances a unified research framework through three objectives: (i) modeling EV proliferation and its value streams to inform grid and market decisions; (ii) defining and operationalizing a Responsible AI (RAI) framework for AVs, with emphasis on bias and fairness risks throughout the AI lifecycle; and (iii) proposing a generative agentic AI-based framework to balance privacy and utility in textual data sharing.

First, a jurisdiction-independent modeling approach is developed to classify factors influencing EV adoption and link them to monetary and non-monetary value streams. This enables grid and market operators to quantify how adoption shifts operational timelines and economic opportunities. Findings highlight the importance of addressing barriers concurrently and demonstrate that coordinated interventions accelerate the transition, while divergent policies can widen cross-jurisdictional disparities.

Second, AVs are examined as socio-technical systems with AI integrated throughout their lifecycle. A comprehensive RAI framework is introduced, covering nine risk domains with actionable mechanisms, and prioritizing bias and fairness. Bias identification and mitigation strategies are demonstrated using public AV datasets.

Third, the thesis proposes a multi-agent architecture for privacy-preserving textual data sharing that combines context-aware utility thresholds, differential privacy for text embeddings, and generative AI for synthetic data augmentation. The evaluation examines utility and semantic fidelity before and after applying privacy mechanisms.

Together, these contributions position next generation mobility as a unified energy mobility data ecosystem, redefining electrification's grid impact, AI's risk and responsibility dynamics, and data sharing's analytics-privacy optimization. Overall, the thesis delivers integrated, actionable, scalable, and trustworthy solutions for electric and autonomous transport.

Dedication

To my grandmother, a highly respected scholar in her era. Your discipline, curiosity, work ethics, and courage set a standard that inspired many who followed, including me.

Though you are not here to see it, your influence remains the driving force behind shaping my values, and for helping me become the person I am today.

This thesis is for you.

Acknowledgments

This thesis is dedicated to the individuals and influences that supported me throughout this journey.

First and foremost, I dedicate this work to my supervisor, Professor Hany E. Z. Farag, whose guidance shaped the direction of this research. I am grateful for your unwavering support, patience throughout each stage, and the professionalism that provided stability during challenging periods. Your humility and depth of knowledge fostered an environment where ideas could develop, questions were encouraged, and I was able to grow with confidence. I sincerely appreciate your mentorship and the example you set for the scholar, collaborator, and individual I aspire to become.

I am deeply grateful to my committee members, Professor John Lam and Associate Professor Manos Papagelis, for their confidence in the research direction and for enhancing this work through thoughtful, constructive feedback. Their guidance at critical milestones, including the thesis proposal, report review, and qualification examination, strengthened both the rigour and clarity of this thesis. Their questions consistently prompted deeper analysis and more comprehensive responses.

I also wish to thank the teachers and professors whose courses during my undergraduate and graduate studies established the foundation for this thesis. Their instruction provided knowledge, frameworks, discipline, and perspective that significantly shaped my academic interests and approach.

To my family, I am grateful for your unwavering strength and support. To my wife, Kadambari, I offer my deepest gratitude. I took away weekends, time, and presence that should have been yours, and you met that sacrifice with faith, encouragement, and calm resolve. Your belief in me carried

me through the days when mine wavered. This thesis belongs to you as much as it does to me.

I also dedicate this work to those who, in both subtle and significant ways, contributed to my personal development. Mentors, friends, colleagues, and loved ones offered perspective, challenged my assumptions, and reinforced the significance of this journey.

Finally, I dedicate this work to the Almighty, the guiding force in nature, who provided strength, curiosity, and clarity at critical moments. Many of the most significant insights emerged during meditation, as challenges were resolved and solutions became apparent beyond conscious effort. I remain profoundly grateful for this grace and the guidance that enabled this journey.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AEGIS	Adaptive Epsilon & Gaussian/Laplace Intelligent Selector agent
AEVM	Available Electric Vehicle Models
AI	Artificial Intelligence
AURA	Adaptive Utility Reasoning Agent
AV	Autonomous Vehicle
AVL	Average Vehicle Life
BEV	Battery Electric Vehicle
CHIMERA	Corpus Hallucination & Iterative Model for Enhanced Resemblance & Anonymity agent
CII	Charging Infrastructure Interoperability
CORTEX	Context-Oriented Reasoning & Threshold-Extraction eXpert agent
CRM	Cost of Raw Material
DER	Distributed Energy Resources
DES	Distributed Energy Systems
DG	Distribution Grid
DGMEV	Distribution Grid Monitoring for EV loads
DI	Disparate Impact
DOD	Depth Of Discharge
DP	Differential Privacy

DPD	Demographic Parity Difference
DSO	Distribution System Operator
ECHO	Experience-Consolidating Heuristic Optimizer agent
EOD	Equalized Odds Difference
ET	Emission Targets
EV	Electric Vehicle(s)
EVCM	Electric Vehicle Charging Management
EVDR	Electric Vehicle Driving Range
EVE	Electric Vehicle Exemptions
EVFMO	Electric Vehicle Fleet Management & Optimization
EVMS	Electric Vehicle Market Share
EVP	Electric Vehicle Privileges
EVSE	Electric Vehicle Supply Equipment
FIA	Fairness in Accuracy
FISSION	Feature Importance & Selection Intelligence ON-agent
FNR	False Negative Rate(s)
FO	Fleet Operators
FPR	False Positive Rate(s)
GAN	Generative Adversarial Network
GHG	Greenhouse Gas
GRNVPY	Growth Rate of New Vehicle buyers Per Year
HOV	High Occupancy Vehicle
IBT	Improvement in Battery Technology
ICV	Internal Combustion-engine Vehicle
IFS	Intersectional Fairness Score
IoT	Internet of Things

LLM	Large Language Model
MAESTRO	Modular Agentic Engine for Strategic Tuning and Reporting Orchestration agent
ML	Machine Learning
OC	Operational Cost
OEM	Original Equipment Manufacturer
ORION	Objective Retention & Information Optimization Nexus agent
P2P	Peer-to-Peer
PC	Purchase Cost
PEM	Parking-lot Energy Management
PHEV	Plug-in Hybrid Electric Vehicle
PII	Personally Identifiable Information
PSO	Parking-lot Schedule Operator
PV	Photovoltaic
QUANTA	Qualitative-qUANtitative Translation Agent
RAI	Responsible Artificial Intelligence
RC	Refueling Convenience
REVPR	Relative EV Proliferation Rate
RIGP	Rate of Increase for Gasoline Price
RIEP	Rate of Increase for Electricity Price
RTCO	Relative Total Cost of Ownership
SCEV	Safety Concerns with Electric Vehicles
SDT	Short-Distance Trips
SENTINEL	Semantic & sENsitivity Text INspector for Embedding aNaLytics agent
TCO	Total Cost of Ownership
TEI	Transactive Energy Initiatives

TNR	True Negative Rate
TOU	Time Of Use
TPR	True Negative Rate
UDA	Unsupervised Domain Adaptation
V2G	Vehicle-to-Grid
V2H	Vehicle-to-Home
V2V	Vehicle-to-Vehicle
VAE	Variational Auto-Encoders
VMPY	Vehicle Mileage Per Year
XAI	eXplainable Artificial Intelligence
ZEV	Zero Emission Vehicle

1 | Introduction, Research Motivation, and Objectives

1.1 Introduction

Transportation and energy systems are undergoing a simultaneous transformation driven by electrification, automation, and extensive digitalization. Traditionally, mobility demand depended on liquid fuels and mechanical control, while electricity systems operated with predictable load profiles and long-term infrastructure planning. This distinction is increasingly blurred. The widespread integration of electric vehicles (EVs) is transferring a significant portion of transportation energy demand to electricity grids. Simultaneously, advancements in autonomous vehicles (AVs) are intensifying the dependence of mobility services on artificial intelligence (AI) for perception, prediction, planning, and control, particularly in scenarios where errors may cause immediate physical harm. Furthermore, the data ecosystem underpinning both electrified and autonomous mobility, including telemetry, sensor data, charging patterns, user preferences, and operational records, has expanded in both scale and sensitivity. Contemporary AI, especially generative and agentic models, increases the economic value that can be extracted from data while also heightening privacy, security, and governance risks associated with cross-organizational data sharing.

This thesis is structured around three research objectives that collectively address critical barriers to scalable, reliable, and trustworthy next-generation mobility. The first objective analyzes EV

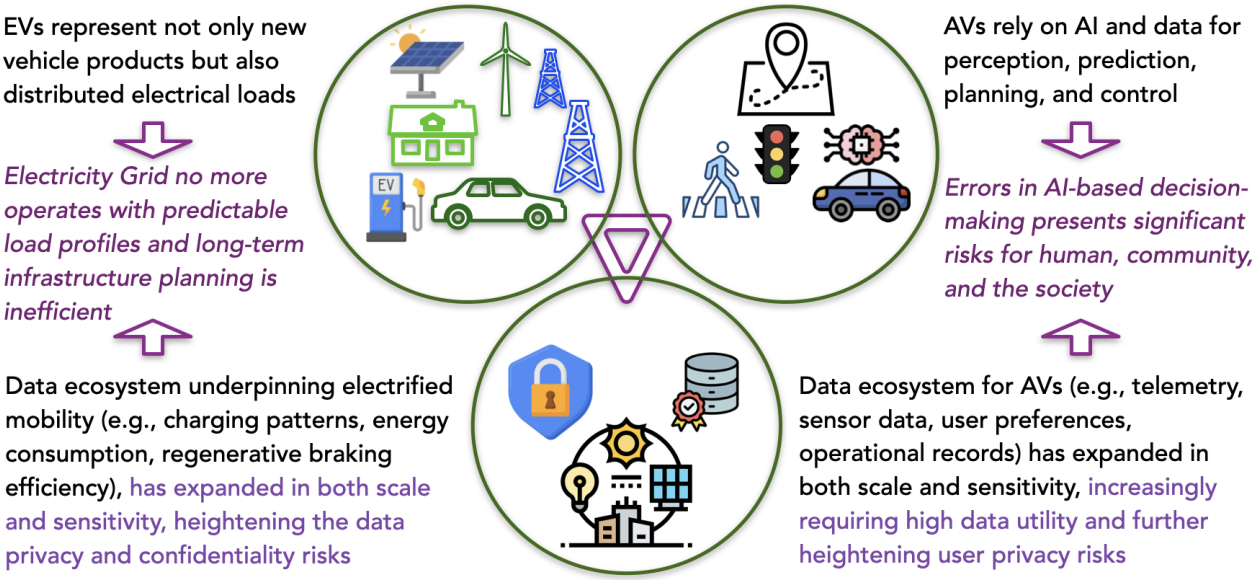


Figure 1.1: Transportation and energy systems are being reshaped at the same time, propelled by electrification, automation, and widespread digitalization.

proliferation, examining the evolution of EV adoption in response to economic, technological, regulatory, and behavioral factors, and how this adoption generates both monetary and non-monetary value streams for stakeholders, including grid operators, market operators, aggregators, fleet owners, businesses, and consumers. The second objective investigates responsible autonomy, specifically how AV developers and regulators can systematically identify, mitigate, and monitor AI risks throughout the lifecycle of AI-enabled AV systems, with a particular emphasis on bias and fairness risks and their practical mitigation. The third objective addresses privacy and data utility in data sharing, assessing how institutions, systems, and devices can exchange textual data while preserving analytical utility and ensuring rigorous privacy protections through an agentic architecture that integrates differential privacy, responsible generative AI, and adaptive user intent modeling.

Although each research objective addresses a distinct aspect of the mobility transformation, namely, infrastructure and adoption dynamics, safety- and fairness-critical AI lifecycle governance, and privacy-preserving data exchange, they share a common systems-level motivation. Electrified and autonomous mobility cannot achieve scale solely through advancements in batteries, sensors,

or machine learning. Effective scaling requires: (i) coordinated incentives and policies that accelerate adoption while maintaining grid reliability and enabling value creation; (ii) comprehensive responsible AI methodologies that align AV performance across all AI risk domains, including but not limited to, safety, fairness, accountability, and legal compliance; and (iii) data governance architectures that facilitate data sharing without compromising privacy or causing downstream harms exacerbated by modern AI.

1.1.1 Electrification as an Energy-Mobility Coupling Mechanism

EVs represent not only new vehicle products but also distributed electrical loads, collectively introducing a novel class of infrastructure and planning challenges. Charging demand is often highly concentrated in specific locations, such as workplaces, multi-unit residences, fleet depots, and fast-charging corridors, and the rate of EV adoption may exceed the pace of traditional grid expansion. The speed of this transition is influenced by the interactions among multiple factors, including economic costs, charging accessibility, technological advancements, policy incentives, and behavioural changes. Simultaneously, the electrification of transportation creates new opportunities. EVs can provide grid services when managed as flexible resources, such as through controlled charging or participation in distributed energy resources (DER) programs. Determining the economic viability of these opportunities requires understanding adoption trajectories and the relationship between proliferation factors and value streams. Accordingly, the first research objective of this thesis frames transportation electrification as both a challenge and an opportunity. If poorly managed, it can strain existing infrastructure. When effectively integrated with grid operations and market design, it can enhance system reliability and create commercial value. In particular, the first research objective is driven by the necessity to forecast EV adoption trajectories and to convert these forecasts into actionable planning timelines and value-stream opportunities for stakeholders.

1.1.2 Autonomy as a Safety- and Society-Critical AI Deployment

AVs centralize AI-driven decision-making within software, models, and their execution pipelines that encompass data collection, training, evaluation, and on-road operation. In contrast to many AI applications, where failures are primarily informational (e.g., poor recommendations, inefficient customer outreach), failures in AV systems are physical and often irreversible. Furthermore, AV deployment alters the accountability framework, as decisions previously made by human drivers are now executed by AI-enabled systems, introducing new ethical, legal, and social acceptability requirements.

In this regard, the second research objective of the thesis addresses these challenges by proposing a Responsible AI (RAI) framework specifically designed for AVs, emphasizing that AI risks encompass not only ethical considerations but also additional AI risk domains, including but not limited to safety hazards, security vulnerabilities, and legal complexities. The framework covers the entire lifecycle of AI intervention and highlights the compounding nature of risks, in which biases introduced during data collection can interact with algorithmic decisions and operational constraints, leading to cumulative effects. Additionally, this research prioritizes practical mechanisms for bias detection and mitigation, including simulations that validate mitigation strategies using publicly available datasets.

1.1.3 Data Sharing, Privacy, and Utility under Modern AI

Electrified and autonomous mobility systems generate and utilize substantial volumes of sensitive data. Sharing this data across institutions, systems, and devices for purposes such as research, planning, regulation, operational coordination, or service optimization can yield significant societal benefits, but it also introduces privacy risks for individuals and organizations. Modern AI fundamentally alters this balance, as devices increasingly collect, process, and utilize user data for diverse applications. Generative AI can synthesize or transform data while preserving semantic

content, and agentic AI can iteratively optimize decisions such as privacy parameter selection. However, these same capabilities may introduce new risks of data leakage and misuse if not properly governed.

Based on the above discussion, the third research objective of the thesis addresses this aforementioned challenge in textual data sharing by proposing an agentic, responsible generative AI-based architecture that jointly optimizes data utility and privacy. This approach formalizes the translation of qualitative user utility expectations, such as perceived completeness or coherence, into quantitative thresholds. It also demonstrates how differential privacy can be applied in embedding space while maintaining semantic fidelity, and how generative augmentation can preserve coherence while minimizing sensitive information leakage. The framework is designed to be adaptive and dataset-dependent, learning from previous iterations and evolving as requirements change.

1.2 Research Motivation

1.2.1 Why EV Proliferation Requires Integrated Modeling and Value-Stream Reasoning

A central challenge in EV proliferation is that adoption is influenced by multiple interacting factors. Technological improvements, such as cost reductions and performance enhancements, may be inadequate if charging infrastructure, consumer economics, or policy incentives are insufficient. Conversely, policy incentives may have a limited impact if technological or infrastructure barriers remain. The first research objective in this thesis is motivated by the need for a comprehensive and flexible modeling approach that can simultaneously incorporate multiple proliferation factors, support scenario analysis across jurisdictions, and quantify the effects of various interventions on EV market share trajectories.

Equally significant, EV proliferation generates value streams that can support, justify, or expedite

infrastructure and market transformations. Grid operators may achieve operational benefits by integrating EVs as DER. Market operators and aggregators can identify new revenue opportunities, while consumers and fleet operators (FO) may benefit from reduced total cost of ownership and innovative service models. A structured mapping of proliferation factors to value streams enables stakeholders to identify where interventions can unlock multiple downstream benefits. This mapping is particularly useful for planning timelines, as it indicates the urgency of grid and market changes required to accommodate growth while capturing value.

1.2.2 Why Responsible AI Must be Lifecycle-Based for AVs

Responsible AI discussions frequently emphasize principles such as fairness, transparency, and safety, but often lack implementable mechanisms spanning the engineering lifecycle. In the context of AVs, this gap is particularly significant. Bias and fairness risks may originate from non-representative data, proxy variables, measurement artifacts, model design decisions, and post-deployment drift; these risks can also compound, resulting in disparate impacts across demographic groups and driving contexts. In this regard, the second research objective is motivated by the need for (i) an integrated perspective on AI risks for AVs across nine AI risk domains, including but not limited to safety, security, ethical, and legal domains, and (ii) a lifecycle framework that delineates when and how risks should be identified and mitigated.

Within this broader agenda, the emphasis on bias and fairness is driven by both societal impact and technical complexity. Bias may be present even when overall accuracy is high, and fairness failures can be localized to specific contexts, such as underrepresented environments or demographic intersections. Accordingly, the second research idea offers detailed techniques for bias detection and mitigation, and demonstrates their effectiveness through simulation studies. This focus on actionable methods seeks to bridge the gap between abstract RAI principles and practical engineering implementation.

1.2.3 Why Privacy and Utility Must be Co-Optimized in Data Sharing

In practical data-sharing scenarios, utility is seldom captured by a single metric, and privacy is rarely a uniform constraint. Different users and tasks prioritize various semantic attributes in textual data, such as relevance or completeness, and datasets differ in their sensitivity and risk of information leakage. Traditional approaches often treat utility modeling, privacy calibration, and user intent as separate issues, limiting adaptability when data characteristics or expectations evolve. The third research objective is motivated by the need for an end-to-end framework capable of (i) interpreting qualitative user intent, (ii) translating intent into quantitative objectives, (iii) dynamically calibrating privacy mechanisms, and (iv) preserving semantic coherence under data privacy constraints.

Furthermore, modern AI introduces dual-use challenges. Generative methods can enhance utility, for example, through data augmentation, but may also inadvertently disclose sensitive structures if not properly managed. Agentic methods can automate decision-making processes, such as privacy budget selection, but may also centralize control, necessitating explicit accountability. Consequently, the third research idea is conceptualized as a responsible generative and agentic architecture, aiming to preserve interpretability, ensure rigorous privacy protection via differential privacy, and maintain measurable semantic utility through embedding-aware evaluation.

1.3 Research Questions

The structure of this thesis is guided by the following fundamental research question and the associated research gaps it identifies:

> *How can electrified and autonomous mobility scale in a way that is simultaneously reliable for energy infrastructure, responsible in AI-driven decision-making, and trustworthy in data sharing through rigorous privacy-utility preservation under the influence of modern AI, i.e., generative and agentic AI?*

This research question is further divided into several sub-questions that are systematically examined in this thesis:

1. How can EV adoption be modelled in a jurisdiction-agnostic manner that accounts for interacting proliferation factors and supports scenario analysis across regulatory regimes and exogenous influences?

2. How can the interdependencies between EV proliferation factors and emerging monetary/non-monetary value streams be structured so that stakeholders can quantify planning timelines and identify coordinated intervention opportunities?

3. What AI risk domains are most salient for AV deployment, and how can these risks be systematically identified and mitigated across the end-to-end AI lifecycle in a unified RAI framework?

4. How can bias and fairness risks in AV systems be detected and mitigated across data collection, algorithm design, deployment, and real-time decision-making, and how do mitigation strategies trade off against performance?

5. How can qualitative user intent about textual data utility be translated into quantitative, context-aware utility thresholds suitable for automated optimization?

6. How can differential privacy and responsible generative AI be combined in an agentic architecture to preserve semantic utility while enforcing privacy constraints, and how can such a system adapt to dataset-dependent characteristics and evolving requirements?

1.4 Research Objectives

To answer these identified research questions, the objective of this thesis is to develop an integrated set of models, frameworks, and methods that enable (i) robust planning and value

realization for EV proliferation, (ii) responsible AI development for AVs through implementable fairness mechanisms, and (iii) adaptive preservation of privacy-utility trade-offs for textual data sharing in the context of modern generative and agentic AI as described hereunder.

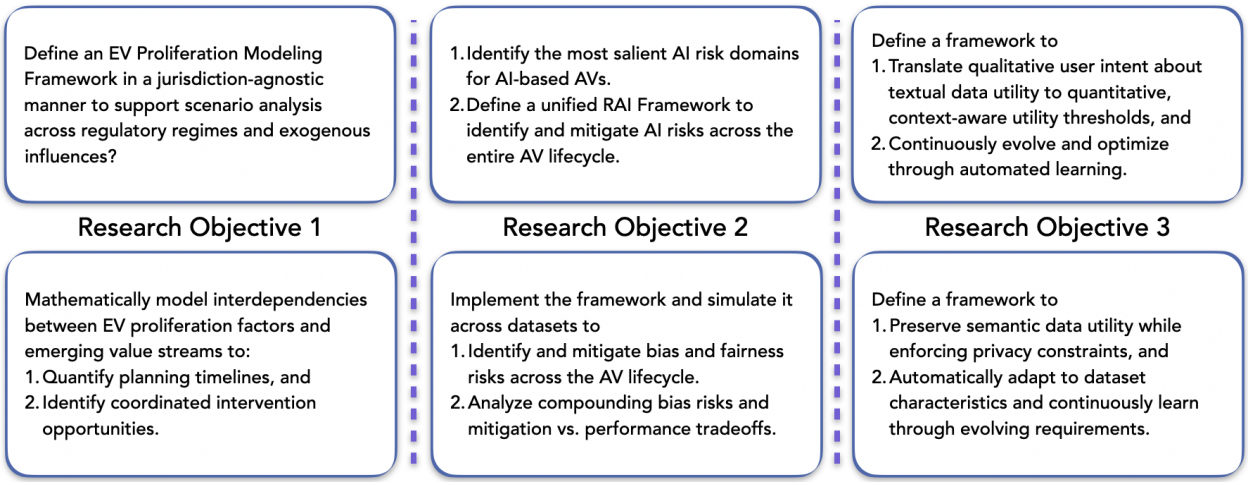


Figure 1.2: Research Objectives of this Thesis.

1.4.1 EV Proliferation and Value Streams

Stakeholders require a robust, scenario-driven understanding of EV market share evolution across jurisdictions and regulatory regimes. However, EV adoption is influenced by multiple interacting factors, and incomplete modeling can lead to misleading planning timelines and underestimation of infrastructure needs. Further, the economic and operational value streams enabled by EV proliferation depend on adoption trajectories and on the relationship between adoption drivers and grid/market opportunities. Therefore, this thesis addresses the problem of creating a comprehensive adoption model and a factor-to-value-stream framework that can support timely investments, grid reliability planning, and data-driven policy decisions across jurisdictions.

1.4.2 Responsible AI Lifecycle for AVs

AV deployment demands systematic management of AI risks across multiple domains, including safety, security, ethics, and law. Existing approaches frequently emphasize ethical principles without providing implementable mechanisms and quantification usable by technologists, and often fail to integrate risk management across the full AI lifecycle. Bias and fairness risks, in particular, can originate at multiple stages and can compound post-deployment, resulting in inequitable and potentially unsafe outcomes. Therefore, this thesis addresses the problem of defining a holistic, lifecycle-based RAI framework for AVs and of demonstrating actionable methods for identifying and mitigating bias and fairness risks.

1.4.3 Agentic Privacy-Utility Preservation for Text Data

Organizations, systems, and devices increasingly share textual data to enable analytics and AI, but this creates privacy risks for individuals and confidentiality risks for institutions. Prior approaches often use fixed utility metrics, static privacy budgets, or narrowly defined objectives, limiting their ability to adapt to heterogeneous datasets and evolving user expectations. Further, integrating generative AI can improve semantic utility but also create new leakage risks, and agentic optimization can improve adaptability but requires transparent control and accountability. Therefore, this thesis addresses the problem of building an end-to-end, interpretable, adaptive framework that jointly optimizes privacy and utility for textual data sharing using differential privacy, responsible generative AI, and agentic learning.

1.5 Thesis Scope and Delimitations

The scope of this thesis is intentionally limited to three primary research objectives and their integration, rather than addressing all aspects of mobility systems, grid engineering, or AI

governance. The first objective focuses on modeling EV proliferation and value-stream mapping as a decision-support framework. The second objective examines Responsible AI for AVs, with particular attention to bias and fairness risks, and simulations that demonstrate mitigation strategies. The third objective investigates the preservation of textual data privacy and utility through a multi-agent architecture that incorporates differential privacy and responsible generative AI.

Broader topics, including commercialization, regulation, or societal adoption, are discussed only to synthesize implications relevant to the three core research objectives, rather than to provide comprehensive coverage. No additional factual claims are introduced beyond those supported by the referenced papers. When citation-supported statements are required, such as in the policy impact context of the first research objective, the thesis relies exclusively on citations from the underlying works.

1.6 Thesis Structure

The thesis is organized into six chapters as outlined below:

- **Chapter 1** presents the research background, motivation, identified gaps, and objectives.
- **Chapter 2** reviews the relevant literature on EV proliferation models, RAI principles, and privacy & utility preserving AI frameworks.
- **Chapter 3** addresses the first research objective by detailing the dynamic modeling of EV proliferation and the value stream framework, simulation results, and associated policy implications.
- **Chapter 4** elaborates on the second research objective by defining a Responsible AI framework for AVs, conducting bias mitigation simulations, and describing implementation strategies.
- **Chapter 5** introduces the agentic AI framework for privacy-utility optimization, describes

the experimental setup, and presents a performance evaluation to address the third research objective.

- **Chapter 6** concludes the thesis by synthesizing findings, discussing cross-domain implications, addressing commercialization challenges, and outlining future research directions.

Integrating the three research objectives provides a comprehensive roadmap for the design, governance, and deployment of intelligent mobility ecosystems that are technologically advanced, ethically aligned, and socially beneficial. The combined focus on responsible AI, agentic governance, and electrification modeling offers a blueprint for achieving transparency, trust, and resilience in future mobility and data infrastructures.

2 | Literature Review

2.1 Electric Vehicle Proliferation

EVs have been gaining traction due to their multiple environmental and societal benefits. One of the most important benefits is the higher energy efficiency (3 to 5 times) over their traditional counterparts, the Internal Combustion-engine Vehicles (ICVs) [1]. Due to zero-emission, Battery EVs (BEVs) produce no tailpipe emissions, and in general, Plug-in Hybrid EVs (PHEVs), which partially run by electricity, also generate lesser tailpipe emissions than ICVs [2]. EVs can help significantly reduce overall Greenhouse Gas (GHG) emissions when combined with the increase in low-carbon electricity generation. Lately, EV charging stations or EV supply equipments (EVSEs) are being deployed worldwide, supporting the case for EVs across multiple transit modes such as shared transportation (buses, taxis) [3, 4], light-motor vehicles (cars, two/three-wheelers) [1], human-operated vehicles (e-Rickshaws) [5], as well as heavy-duty vehicles for short-range urban deliveries [1, 6]. Furthermore, EV manufacturers worldwide have increased their investments, resulting in various EV models offered today, offering broader choices to consumers across various segments. Notwithstanding, favorable and effective policies are crucial to faster EV proliferation by lowering the upfront investment cost gap, promoting charging infrastructure, and ensuring a smooth integration of EV charging demands into power systems and the overall grid [1].

Overall, the proliferation of DER, such as solar photovoltaic (PV), battery storage, and EVs, is an emerging challenge to the power grid. However, unlike other DER, EVs cause more significant

challenges [7] due to being mobile and mainly creating clustered density of load at public charging, workplace charging, and multi-unit residential charging facilities [8]. Moreover, as commercial EV fleets emerge, the charging facilities will likely be concentrated at seaports, airports, and other retail and wholesale transportation distribution centers, further increasing the problems emerging due to load density. Recent trends indicate an increased focus on installing fast or higher-speed charging stations to promote EV proliferation, which may pose an increasingly severe threat to grid stability if not planned carefully [1, 9]. Accelerated EV deployment may soon exceed the capacity of existing electrical infrastructure, which has already started showing early signs of weaknesses [7, 10]. A typical infrastructure upgrade or new capacity deployment takes many months to a few years due to the time-consuming processes involving planning, regulatory approvals, project commissioning, and execution. Therefore, consideration of opportunities to mitigate costs and incorporate load management strategies to minimize on-peak charging is the need of the hour.

While the rapid proliferation of EVs challenges the grid's reliability, it allows for a bottom-up approach where EVs enable new energy services to allow consumers to buy and sell energy. Connected devices now enable communication and control, which was impossible in the past. To some extent, advancements in technology, including two-way near real-time communication and the solid-state power conversion electronics of EVSE or EV chargers, are together helping mitigate the potential impact of new loads and provide valuable grid services. Utilities have already started to leverage this as an opportunity to support essential grid services such as congestion management and voltage regulation [1, 11].

Managed EV charging, Vehicle-to-Grid (V2G), and Vehicle-to-Home (V2H) services have started helping solve issues presented by EVs at the distribution network level. At the same time, EV charging facilities have started installing solar PV and energy storage systems, further enhancing their capability to support grid operators [2]. In addition, market operators, utilities, and aggregators have started creating new value streams leveraging the consolidated capacities of controllable and dispatchable EV resources. A value stream can be an initiative generating monetary, non-monetary,

or both types of value to one or more beneficiaries, such as system operators, utilities, aggregators, end-consumers, and the environment.

2.2 Factors Impacting EV Proliferation

Several studies have analyzed the factors behind EV proliferation and have also forecasted growth patterns based on those factors. Based on the studies, public and private sectors industry initiatives, as well as regulatory directions, each factor can be classified under different categories [12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. The first category covers technological factors accounting to those which are related to battery, EV raw material, or overall EV advancements. The second category covers jurisdictional policies typically directed toward supporting environmental and sustainability goals. On the other hand, economic factors are direct drivers of EV purchase decision making, including the purchase cost, purchase tax, and relative operational cost of EVs versus ICVs. Lastly, some new factors have recently emerged, such as due to the COVID-19 pandemic, and are covered in the others category. The following sections describe each category and associated factors.

2.2.1 Technological Factors

Among the various technological factors influencing a vehicle buyer's decision to purchase an EV, the availability of EV charging infrastructure, predominantly driven by public charging stations across most geographies, is the most prominent influencer [12, 16, 18]. In countries such as India, with currently small EV charging networks, this is a significant adverse influence, and the government is investing in deploying more EVSEs [22]. Despite the government's efforts to define EV-friendly policies, Ref. [23] highlighted the need for more aggressive policy mandates by reducing taxes on EVs (BEV, PHEV) while simultaneously imposing higher taxes on ICVs. Furthermore, EV proliferation in geographies like India faces unique challenges, such as a limited local supply chain for powertrain and battery pack assembly, leading to high EV and EVSE prices, as

well as high real-estate acquisition costs to install EVSE, resulting in insufficient EVSE availability, as discussed by ref. [24]. On the contrary, EV infrastructure has been growing rapidly in the USA and China for several years [9, 12, 22]. The number of EVSEs in the USA has grown from 75,000 to 100,000 between 2020 and 2021, while in China it is 210,000 [25]. However, the number of gasoline stations in the USA has only grown from 111,100 in 2016 [26] to 115,000 in 2020 [27]. These trends highlight strong support for EV proliferation in some geographies while discouraging the use of ICVs.

Lately, the availability of fast-charging technology has been gaining traction. As per a recent report [9] from the National Renewable Energy Laboratory (NREL), about 12.0% and 49.6% of the Level 2 and DC fast EVSE, respectively, required to meet projected demand in 2030, have been installed as of Q1 2020 in USA. Moreover, there are 13,627 public and workplace DC fast-charging EVSEs and 71,975 public and workplace Level-2 EVSEs available in the United States. However, a similar EV fast charger proliferation level is required to help EVs penetrate globally. In addition to fast chargers, improvements in battery technology have collectively contributed to a significant reduction in time to charge an EV fully. Recently, StoreDot [28], a battery manufacturing firm, has released mass production-ready batteries that can fully charge in around five minutes and has successfully demonstrated it for smaller batteries in phones, drones, and electric scooters [29]. The firm highlighted that using silicon in place of graphite for battery electrodes has primarily led to this development and helped bring the cost closer to existing Li-Ion batteries, further helping obtain investments from major automotive manufacturers globally.

EV infrastructure interoperability challenges play a significant role when an EV buyer cannot charge the vehicle using an EVSE from another make when needed. Some EVSE manufacturers focus on interoperability [30, 31] while some [32] do not, consequently impacting EV growth. According to a recent Natural Resources Defense Council report [22], the EV charging infrastructure interoperability is one of the most critical factors towards EV proliferation and is increasingly more significant in large markets such as the USA, China, and India. Another critical factor influencing

EV buyer's decisions is EV driving range on a full charge [1, 12]. While EV manufacturers are working on improving the driving range, Ref. [33] has provided a unique model for predicting EV driving range under the influence of factors such as days, temperature, and the depth of discharge (DOD) of a battery pack.

EV proliferation is continuously being impacted by the recent incidents of safety concerns around EVs [34], primarily due to batteries catching fire caused by overheating or poor health. Notable recent efforts to help address these issues include identifying determinants impacting battery health [35] and the creation of tools to measure battery health over time [36]. While recent advancements in battery technology have minimized these risks, it is vital to address EV safety issues in totality to sustain EV growth. Efforts by Tesla and Volkswagen have proven to be successful in recent years [37, 38, 39]. EVs appeared as catalysts in the overall industrial development due to them being the potential enabler of cost reduction in battery and copper-based technologies [2]. Battery cost, typically accounting for up to 30% of EV cost [40], has fallen 87% since 2010 and is expected to drop another 60% by 2030 [41]. Furthermore, demand for copper is forecasted to rise nine-fold by 2027, given that it is the second most crucial component constituting the majority of equipment costs in EVs and EVSEs [42]. Consequently, similar to batteries, copper prices need to drop in coming years, as highlighted by global auto manufacturers [43].

2.2.2 Jurisdictional Policies to Support Environmental Goals

Jurisdictional directions significantly impact a vehicle buyer's decision to purchase an EV. In the last few years, governments across the globe have started intervening in setting the targets for GHG emissions, forcing vehicle manufacturers to shift their strategy towards making more EVs over ICVs, and avoid fines with reputational setbacks [1, 22, 44]. One such example is the European Union's regulation on CO_2 emission performance standards for new passenger cars and vans [45], which could bring penalties for up to 50% of vehicle manufacturers [12]. GHG emission targets across jurisdictions could vary, where some have been more stringent over others, as highlighted by

online resources such as the climate action tracker [46]. In the USA, the Center for Climate and Energy Solutions [47] provides a granular view of GHG emissions across states to help establish GHG mandates.

Another recent trend is imposing fines, punitive taxes, or complete bans on older ICVs to address air pollution concerns [48]. Similarly, some jurisdictions have started providing EV-related privileges (EVP) and exemptions (EVE) to encourage EV adoption. Examples of privileges may include dedicated EV lanes and allowing high-occupancy vehicle (HOV) lanes for EVs with one occupant [49, 50], unlike the general rules requiring vehicles to carry at least two passengers. Ref. [58] highlighted the benefits of exemptions to promote customer adoption of EVs. Furthermore, introducing self-driving features is increasingly encouraging EV adoption, where emerging ideas suggest dedicated lanes for autonomous cars, especially at toll booths [51]. Recent studies point towards this direction, where Ref. [52] studied the impact of dedicated lanes for AVs on traffic flow throughput, while Ref. [53] presented a conceptual framework to design and operate dedicated lanes for connected and automated vehicles on motorways. As of November 2020, at least 45 states in the USA, along with the District of Columbia, offered incentives to support the deployment of EVs or alternative fuel vehicles and supporting infrastructure, either through state legislation or private utility incentives [54]. Legislative incentives include measures that provide HOV lane exemptions, financial incentives for purchasing EVs and EVSEs, exemptions from vehicle inspections or emissions testing, parking incentives, and utility rate reductions.

In countries such as Japan, the lower availability of EV models over ICVs is one of the unique factors that is negatively impacting EV growth [1]. While vehicle manufacturers globally have been investing in all types of EVs, including PHEVs and BEVs, Japan's unique focus thus far towards self-charging hybrid EVs has slowed the growth of overall EV proliferation [55]. It is mainly because the country's largest manufacturers have been left behind in the latest technological shift in the automobile sector while keeping their focus toward self-charging hybrid vehicles [56]. However, recently, Japan rolled out policies for vehicles and chargers to achieve its target of "next-generation

vehicles" sales to account for 50–70% of the total vehicles in the country by 2030 [1, 57].

2.2.3 Economic Factors and Related Policies

Multiple economic factors play an essential role in decision making while buying a vehicle [1, 16, 58]. Certain municipalities across the globe have launched programs to reduce costs for vehicle charging in public charging stations and, in some cases, eliminate parking charges for EVs [1, 54]. Financial incentives in the form of cash subsidies and reduced insurance on EVs have encouraged EV adoption, while certain jurisdictions have maintained or increased taxes on ICVs to encourage buyers to own an EV. Ref. [58] demonstrated the importance of vehicle purchase tax and carbon tax for EV adoption in the short and long terms. Most people look for an overall lower cost of ownership of vehicles, which includes one-time purchase cost, tax, and the operational cost of the vehicle, covering cost per kilometer (or mile) and maintenance cost over the vehicle lifespan. Ref. [16] proposed a dynamic model of EV adoption, which helps calculate the overall life-cycle costs of EVs and ICVs as a result of operational and one-time costs for each.

Multiple studies have assessed the impact of policies on EV market share where a unique model presented by Ref. [59] predicts PHEV market share under alternative policy settings. Different subsidy options are analyzed for high- versus low-income consumers to examine their impact and help identify strategies to maximize EV proliferation. Further, a different study [60], argued that PHEV can be made more economical by employing focused incentives based on factors such as household income, vehicle disposal, geography, and vehicle travel usage. In addition, the study challenges the existing policies, such as in the USA, which offer more significant subsidies for PHEVs with larger batteries, and argued that offering similar incentives regardless of the battery size could result in higher EV proliferation. Ref. [61] adopted a more practical approach to conduct an econometric study of purchase incentives by analyzing actual data on combined BEV and PHEV sales from 32 European countries between 2010 and 2017. Their study concluded that factors such as household incomes, fuel prices, and supporting financial incentives impact BEV/PHEV sales and

thus can facilitate their diffusion. A similar study [62] examined the impact of purchase incentive policies on EV proliferation and found that up to 35% of EV sales could be attributed to the purchase incentives.

2.2.4 Other Factors

EV transition is expected to be slower in nations with lower per capita income where high population and cultural differences regarding mobility models could be additional factors impacting transition to EVs [12, 22]. For example, India, dominated by mass and low-cost mobility models, is a region that EV manufacturers have not penetrated so far because of comparatively higher EV prices than ICVs. However, fleet owners globally, including the developing nations, have been investing in buying EVs due to lower operating costs and leveraging government incentives [1, 63]. Fleet sales represent a significant proportion of all cars sold globally and are an essential driver for overall EV sales [13].

In light of COVID-19, investments in fleets were initially stalled for a few months as corporates reduced their expenditure and prioritized other investments [1]. However, fleet investments started rising again as the world is approaching normal, and the forecasts show it in favor of EVs [64, 65]. Moreover, trends show an overall decline in sales in the automotive industry, where, in some geographies, close to 50% of prospective vehicle buyers now plan to keep their existing vehicles for longer than initially intended [1]. However, the same reports highlighted comparatively less impact on EV sales than on overall automotive sales. In addition, since the COVID-19 pandemic has hit the world, people's driving behaviors have changed globally [66]. A significant population requires short-distance trips to a grocery store, taking away the current concerns for EV owners, range anxiety, and long wait times to charge their vehicle [12]. Moreover, people prefer charging their vehicles at home versus going to a gas station to refill the vehicle due to pandemic-related safety concerns. Ref. [67] studied this aspect in detail and provided evidence of significantly different post-pandemic travel behavior compared to pre-pandemic. Overall, these factors collectively influence a vehicle

buyer's decision to favor EVs over ICVs.

Lately, transactive energy initiatives have been gaining widespread adoption, driven by both jurisdictional mandates and public-private partnerships [68]. Multiple pilot projects [69, 70, 71] have demonstrated that transactive energy systems leveraging EVs can generate monetary value through revenue and infrastructure deferral. Simultaneously, they can provide non-monetary benefits such as increasing grid reliability and environmental and social benefits. Concurrently, multiple academic studies [72, 73, 74] understood the aspects of transactive energy and proposed models to understand and demonstrate their role in supporting electricity networks. Ref. [75] highlighted the importance of EVs' transaction behaviour and their interactions with buildings in establishing a sustainable transactive energy community from physical energy space, data cyberspace, and human social space perspectives. Others [76, 77] have proposed models to leverage EVs for energy participation by adopting transactive energy concepts. Overall, these approaches have successfully demonstrated the importance of DER, specifically EVs, in supporting the grid and generating new value opportunities, as discussed in the upcoming sections.

2.3 Value Streams within the EV Domain

Recent studies [78, 79, 80, 81], pilot initiatives [69, 70, 71], and industry solutions [32, 82] have demonstrated the capabilities of EV chargers. They have also highlighted the exponential increase in EV proliferation and indicated that the value obtained from the EV-based ecosystem solutions would increase at a similar or even higher rate. This section studies EV proliferation's impact on the utility grid and value streams arising in this domain. It further examines the EV proliferation from an economic perspective where multiple ecosystem solutions, such as in the space of EV charging management, and parking lot energy management, could generate new monetary opportunities for existing and new businesses. In addition, large organizations relying on vehicle fleets for their day-to-day operations could increasingly benefit from EVs due to their significantly

lower operational cost than ICVs. Lastly, household consumers could save a lot of dollars on rising fuel prices by steering toward more economical EV commute and ownership options.

2.3.1 EV Charging Management (EVCM)

A rapid increase in EV adoption is leading to peak-demand problems for the grid [83, 84], which are becoming more common in recent years, where the clustered density of EVs is further aggravating those issues [8]. Multiple studies [80, 81] demonstrated the benefits of managed EV charging approaches to address peak-demand-related issues encountered at the transformer and feeder level. If not managed carefully, these issues at scale can increase grid congestion beyond the feeder level up to the substations, leading to network instability, infrastructure/equipment failures, and outages.

An additional strategy by utilities to manage grid congestion and loading issues is by encouraging DER owners to participate in time-of-use (TOU) pricing programs. DER owners could be incentivized for their active participation by allowing utilities to control their EV charging rate to promote participation in these programs. Pilot programs like the one by Colorado in partnership with Xcel Energy [69] and the other by Chicago with ComEd [70] are good examples of DER participation. Similarly, studies [71, 78, 79], are selected examples of EV energy participation in the last decade. TOU pricing programs offered by utilities to EV owners can provide significant monetary benefits in the long run by offering a suite of services such as registration of EV assets, operational and compliance management, and data analytics, thus encouraging participation by EV owners [69, 70, 71, 78, 79].

2.3.2 EV Fleet Management and Optimization (EVFMO)

A more recent group of beneficiaries in the EV space are the FO managing EV charging for themselves, or on behalf of fleet owners, by optimizing charging schedules for the monetary, grid, and environmental benefits. EV charging schedule optimization can be performed in multiple ways, where the rudimentary strategy is to perform it without grid participation. It involves generating an

EV charging schedule for the FO without coordination with the network or market operator. As described by [80, 85], this approach focuses on maximizing FO profits by predicting the energy requirements of their fleet without considering load impacts on the grid.

While the above strategy does not help resolve grid congestion, a network/market operator-coordinated EVFMO approach can do so. Refs. [80, 85] have highlighted the benefits of this approach where it could be further implemented as a planned or an online coordinated system. A planned coordination approach involves generating charging schedules based on network information sharing at specific times of the day. It helps resolve congestion during peak hours and optimizes well between FO profits and minimizing Distribution Grid (DG) congestion. Despite having no real-time communication between the FOs and the DG for congestion management, Distribution System Operators (DSOs) can override the charging schedule to mitigate network congestion. However, a more sophisticated online coordinated market-based approach involves scheduling in real-time with the network/market operators. In this case, the charging schedule is based on the network state at any given point in time. Moreover, the network operator can override the charging schedule to manage grid stability, which FOs can further leverage to improve their predictive modeling for scheduling optimization.

2.3.3 Vehicle-to-Grid/Home/Vehicle (V2G/H/V)

Multiple studies [86, 87] have highlighted the benefits of using energy stored in an EV battery for participation in demand response, energy arbitrage, or vehicle-to-grid (V2G) services. Although this approach has not yet seen widespread adoption, it helps enhance the value obtained from an EV when it is not on-road but connected to a bi-directional EV charger. Similarly, Refs. [86, 87] have highlighted the benefits of vehicle-to-home (V2H) techniques where energy stored in EV batteries can be used to fulfill load requirements within a residential or an industrial premise. However, similar to V2G, this approach can rapidly deteriorate EV battery lifespan due to multiple charge/discharge cycles throughout the day and is recommended only for emergencies. Leveraging the concept of

V2G/H, peer-to-peer (P2P), or vehicle-to-vehicle (V2V) EV charging has recently been seen as an appealing business model while the associated infrastructure is maturing. Studies [85, 88, 89] have presented two-way energy trading scenarios where bi-directional EV charging infrastructure can significantly eliminate concerns about EV range.

Currently, a typical bidirectional EV charger is more than six times as expensive as its unidirectional counterpart [90]. Automakers such as Tesla [91] and Volkswagen [92], as well as third-party companies such as Quasar [93], are actively working to reduce these costs and facilitate large-scale adoption. In contrast, existing industry regulations that do not support the standardization of bidirectional EV charging stations hinder their widespread deployment and limit their viability as a business model [92]. At present, growth opportunities in V2G or V2H applications remain limited. Nevertheless, the widespread adoption of bidirectional EV charging is expected to create additional opportunities for P2P or V2V charging, potentially enabling a broader range of real-world applications.

2.3.4 Parking Lot Energy Management (PEM)

This approach involves real-time scheduling of EV charging stations, in which the parking-lot schedule operator (PSO) collaborates with distribution system operators (DSOs) who may override charging schedules to maintain grid stability. PSOs must employ real-time scheduling and adapt charging schemes to prevent energy imbalances while participating in the power market. Access to current network congestion data enables PSOs to conduct real-time risk analysis and establish effective charging schedules. Typically, EVs follow the prescribed schedule; however, if the grid's technical operations are compromised, DSO intervention, such as load shedding, can supersede PSO management. Unlike the online approach for EV fleet charging optimization described by [80], the online approach for parking lots is more challenging due to limited or no authority over the charging patterns of incoming vehicles [94]. Various techniques have been proposed, with some focusing on non-utility coordinated offline methods [95, 96, 97, 98], while others [94, 96, 99] have developed

more advanced grid-coordinated strategies.

2.3.5 DG Monitoring for EV Loads (DGMEV)

This approach requires analyzing EV load patterns within the distribution network to characterize charging behaviour, thereby supporting both short- and long-term distribution network planning and enabling more accurate risk analysis. Implementing this approach requires near-real-time communication with the utility control room and closer integration with planning and risk analysis processes. Furthermore, as DER penetration increases, seamless integration of control room technologies with DER management solutions is essential to obtain real-time network state information and manage the distribution network effectively. To date, research has primarily focused on understanding EV charging behaviour, as documented in various studies [100, 101, 102] and reports [103, 104]. However, the literature review indicates that no large-scale, real-world implementations exist in which utilities have integrated this approach into distribution networks.

2.4 Responsible Artificial Intelligence for AVs

The development and implementation of AI in AVs has generated significant interest and concern regarding the ethical implications of this technology [105]. In recent years, the discourse on responsible and ethical AI in AVs has become increasingly prominent [106]. Existing research underscores that ethical considerations are essential for user acceptance of AV technology [107, 108]. The transition from human-driven vehicles to AI-powered AVs represents a significant transformation in transportation, prompting extensive discussion of the associated moral dimensions [109]. Ethical standards and moral compliance in AI systems that operate AVs have emerged as critical concerns for industry stakeholders, who are actively debating the social, economic, and quality-related impacts of the technology [110, 111]. As apprehensions regarding the unintended consequences of AI increase, various frameworks have been proposed to incorporate ethical analysis into engineering practices

[112]. Researchers are also investigating the application of deep learning models to simulate moral decision-making in AVs, thereby enabling these systems to address ethical dilemmas [113]. This ongoing discourse underscores the need to address ethical issues in the development and deployment of AVs. Effective management of AI-related risks, including safety concerns, will be essential for achieving social acceptability and widespread adoption of AV technology [114].

RAI extends beyond addressing ethical risks associated with AI; it encompasses the development and use of AI systems with consideration for their societal impact, aiming for safe, ethical, and trustworthy outcomes [115]. As RAI moves from theoretical principles to practical implementation, there is a growing demand for comprehensive, actionable RAI frameworks. Ref. [112] introduces two such frameworks: one emphasizes a responsible design process that integrates ethical and well-being considerations throughout all stages of AI development, including data analysis, ideation, and prototype evaluation; the other addresses the impacts of AI both before and after deployment. Another study [116] reviewed 84 ethical guidelines and identified 11 core principles: transparency, justice and fairness, nonmaleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity. Ref. [117] underscores the importance of leveraging ML and mathematical optimization, including Quantum AI, to enhance vehicle sustainability, a key aspect of RAI. Similarly, [118] emphasizes the importance of sustainability in the transition from EVs to autonomous EVs. The study by [119] examines specific ethical considerations such as safety and privacy in AI-driven AVs, focusing on dilemmas related to user consent, data collection, and algorithmic bias. Ref. [120] applies Multi-Criteria Decision Analysis (MCDA) to develop formal questionnaires for assessing the social and ethical impacts of AI in autonomous systems, with the objective of establishing a rational and ethical sociotechnical system for AVs. Although the survey size of 19 respondents is insufficient for robust conclusions, the analysis provides qualitative insights with limited quantification for AI system development. Similarly, Ref. [121] identifies key facilitators and barriers, such as autonomy, privacy, liability, security, data protection, and safety, influencing self-driving vehicle adoption, but does not provide quantitative measures to support

practical AI adoption. In contrast, Ref. [122] presents a literature survey of ethical issues related to self-driving cars. Overall, the current literature lacks both comprehensive risk quantification and coverage of all potential risks within a single study, which this thesis aims to address.

While most existing studies address only ethical components and do not encompass the full scope of RAI, some works introduce elements such as justice and solidarity that extend beyond technological considerations but lack practical mechanisms for implementation in AI-based systems. Additionally, no current approaches provide quantification of RAI components that technologists can utilize during the design and development of AVs.

2.5 AI-Risks for AI-based Autonomous Vehicles

This section provides a comprehensive list of risks, organized according to nine risk domains relevant to AI-based AVs. The identified risks are derived from a review of academic and industrial literature on the AV industry, as well as from other domains applicable to AI-based AVs. The scope of this research is limited to risks emerging specifically from AI, excluding those associated with mechanical operation, manufacturing processes, or other unrelated risk categories.

1) Safety Risks: Unpredictable or unreliable behaviour resulting from malfunctions in AI systems can cause unexpected outcomes or loss of control, creating safety hazards for passengers and other road users [123]. AVs operate in environments shared with human-driven vehicles, pedestrians, and cyclists, all of which may behave unpredictably, thereby increasing the risk of accidents if not effectively managed. Furthermore, AVs may encounter situations that require rapid decision-making, which can further elevate accident risk. Safety standards such as ISO 26262 [124] address the functional safety of electrical and electronic systems in vehicles by defining risk assessment processes and safety integrity levels to ensure that system failures, including sensor malfunctions, software bugs, or hardware faults, do not result in hazardous situations. ISO 21448 [125] similarly addresses hazards arising from system limitations or foreseeable misuse, even when all components function as

intended. This standard emphasizes the identification and mitigation of risks due to environmental complexity or system design limitations. Aligning risk mitigation strategies with standards such as ISO 26262 and ISO 21448 enables AVs to proactively address safety risks, thereby enhancing public trust and regulatory compliance.

2) Security Risks: AVs, like other software-driven technologies, are vulnerable to hacking, which introduces risks such as vehicle hijacking, unauthorized access to sensitive data, and manipulation of sensor inputs [126]. These breaches threaten both individual safety and privacy. For example, hackers may exploit system vulnerabilities to remotely control vehicles, creating hazardous road conditions. Unauthorized access to AV data can result in the theft of personal or financial information. Furthermore, malicious alteration of sensor data may cause AVs to make erroneous decisions, endangering passengers and other road users. Consequently, robust cybersecurity measures are critical to maintaining the integrity and security of AVs and the data they process.

3) Privacy Risks: The collection and storage of sensitive data in AVs, such as location, biometric information, passenger behaviours, and preferences, raise significant privacy concerns [127]. Unauthorized access to this data may result in privacy violations, unwarranted surveillance, or identity theft. Biometric data, which uniquely identifies individuals, is particularly sensitive and, if misused, can facilitate identity theft or unauthorized profiling. Additionally, data on passenger behaviours and preferences can be exploited to manipulate consumer actions, target individuals with personalized advertisements, or enable intrusive surveillance. Protecting the confidentiality and integrity of such data is essential to uphold privacy rights and prevent exploitation.

4) Ethical and Moral Risks: AVs face ethical dilemmas when required to choose between prioritizing the safety of their occupants or that of other individuals on the road [128]. These scenarios raise critical questions regarding moral accountability and societal norms, highlighting the ethical responsibilities embedded in AV technology. The complexity intensifies when AVs must make rapid decisions that affect the well-being of occupants, pedestrians, cyclists, or other drivers. For example, in situations where an AV must choose between swerving to avoid a pedestrian or maintaining its

course to protect passengers, core ethical principles are implicated. This challenge illustrates the need to balance individual safety with minimizing harm to others, requiring careful consideration of societal values and ethical frameworks. Resolving such dilemmas extends beyond technical solutions and necessitates broader societal dialogue and regulatory development. Addressing these issues requires interdisciplinary collaboration among experts in philosophy, ethics, law, and technology.

5) Bias and Fairness Risks: AI algorithms in AVs may perpetuate or intensify societal biases, leading to unfair treatment or discrimination based on race, gender, or socioeconomic status [129, 130]. If training data contain biases or inaccuracies, AI models can replicate these issues, resulting in discriminatory outcomes. Such biases may manifest as unequal treatment of marginalized groups or increased disparities in access to transportation. The broader societal impact includes the perpetuation of systemic inequalities and the hindrance of efforts to achieve a more inclusive and equitable society.

6) Legal and Regulatory Risks: Assigning accountability for AV-related accidents is complex, involving questions of legal responsibility, insurance, and liability distribution among manufacturers, operators, and users [131]. The unique challenges of autonomous driving require both the interpretation of existing legal frameworks and the development of new regulations. Implementing traceable audit trails for AI decisions can enhance accountability and transparency by logging actions and decision rationales, facilitating root cause analysis after incidents. Automated compliance checks can ensure AVs adhere to evolving legal requirements, such as stricter pedestrian detection rules. As AVs operate across diverse legal jurisdictions, establishing consistent and adaptable regulatory standards is essential for clarity in legal proceedings and liability attribution [132]. Continuous adaptation of legal systems is necessary to keep pace with technological and ethical developments, and regulatory bodies must proactively update regulations to balance innovation with public safety.

7) User Trust and Acceptance Risks: Incidents involving AVs, such as accidents or privacy breaches, can erode public trust and hinder widespread adoption [127]. Limited user understanding of AV capabilities and limitations may lead to overreliance on automation or misconceptions

about system performance, potentially resulting in unsafe practices. Comprehensive education and transparent communication about AV functionalities are essential to foster informed users, promote safer interactions, and build greater trust in autonomous driving technologies.

8) Societal and Economic Risks: Widespread adoption of AVs may lead to workforce displacement in sectors such as transportation and logistics, raising concerns about unemployment and socioeconomic inequality [133]. The proliferation of AVs could also significantly impact urban infrastructure, land use, and travel patterns, highlighting the need for careful urban planning to address congestion, environmental effects, and equitable access to mobility [133, 134]. Policymakers, urban planners, and stakeholders must collaborate to develop strategies that balance the benefits of AVs with the need to promote societal well-being and sustainable, accessible transportation.

9) Sustainability Risks: The production and maintenance of AVs require significant energy and resources, with sensor and computing hardware manufacturing contributing to carbon emissions and resource depletion [133]. Increased convenience may lead to higher travel demand, resulting in more vehicle miles, greenhouse gas emissions, and congestion [134]. Infrastructure development for AV deployment, such as roads and charging stations, can exacerbate environmental issues including habitat loss and land use changes. Electric AVs, dependent on battery technology, raise concerns related to resource extraction, energy-intensive manufacturing, and hazardous waste, while nonelectric AVs may further contribute to energy consumption and environmental degradation.

2.6 Optimizing Data Utility and Privacy

Data collected by organizations and agencies is essential in the contemporary information age, as it provides significant value for reporting and analytical purposes. However, the disclosure of such data and the information derived from it presents substantial risks to individual privacy. Additionally, data disclosure can compromise confidential information about organizational assets and infrastructure, such as transformer and network relay locations within electric utility systems.

When data is shared between institutions on behalf of individuals or organizations, a central challenge is to balance privacy and utility, and to maximize both [135, 136, 137, 138]. Several key questions emerge: What are the appropriate definitions and boundaries of data privacy and data utility from the perspective of data-exchanging parties? Given a specific privacy definition and utility requirement for a dataset, what strategies should be employed to maximize both privacy and utility, or to achieve a specified target for a particular task? Furthermore, given a dataset's inherent utility, are there methods to enhance it while preserving its essential characteristics and protecting privacy?

Differential privacy (DP) constitutes a rigorous, mathematically grounded framework that enables meaningful data analysis while protecting individual confidentiality. Introduced by Dwork [141], DP ensures that the output of a randomized algorithm is nearly indistinguishable regardless of whether any single individual's data is included. This guarantee is formalized through the ϵ- and (ϵ, δ)-DP definitions, which bound the extent to which a query result can change between neighbouring datasets. Dwork and Roth [140] further developed the algorithmic foundations of DP, introducing mechanisms such as the Laplace, Gaussian, and exponential mechanisms, as well as concepts including global sensitivity, composition theorems, and privacy amplification. These tools provide a comprehensive framework for designing private algorithms with quantifiable trade-offs in accuracy. Li et al. [139] expanded this theoretical foundation to practical applications, describing methods for deploying DP in real systems, selecting noise parameters, managing privacy budgets, and balancing utility with privacy across interactive, non-interactive, centralized, and local settings. Collectively, these contributions establish DP as a unified paradigm bridging theory and practice in data privacy preservation.

In the electric-utility sector, privacy-preserving aggregation techniques have become a cornerstone for real-time monitoring of consumer electricity usage while safeguarding individual data confidentiality within smart grid communications [142, 143, 144, 145, 146, 147, 148, 149]. For example, ref. [142] introduced a protocol that combines cryptographic safeguards with efficient data aggregation. However, this work primarily addresses encryption and aggregation mechanics, without

considering personalized agentic control or textual privacy concerns. Subsequent studies have extended this work to household-level data collected via smart meters and other Internet of Things (IoT) devices. Notable frameworks include those proposed by [143, 144, 145], while comprehensive surveys on IoT privacy are provided in [146, 147]. Ref. [148] presented a DP framework for blockchain-integrated IoT systems, introducing a mathematical model that dynamically tunes privacy budgets based on transaction frequency and trust levels among IoT nodes. The primary contribution of this study is the integration of DP mechanisms into distributed architectures, which ensures end-to-end confidentiality during data sharing while maintaining operational throughput. Despite these advancements, the approach remains focused on architecture and mechanisms, lacking agentic intelligence capable of adaptively learning or translating qualitative preferences into quantitative metrics. Specifically, the framework by [147] employs federated learning to iteratively train predictive models under local differential privacy; the privacy–utility tradeoff itself is not learned or adapted. Privacy parameters and noise mechanisms are predefined, and no feedback-driven or agentic mechanism exists to evolve privacy thresholds, utility metrics, or user intent representations over time. In addition, ref. [149] investigated edge-computing architectures designed to mitigate cybersecurity and privacy risks by establishing secure processing layers for physical systems and their associated data assets. Although many of these contributions employ DP concepts and often achieve superior privacy-utility trade-offs compared to non-DP baselines, they do not address all limitations. A persistent limitation is the absence of intelligence and adaptive contextual reasoning, which would enable translating qualitative user perceptions into quantitative utility metrics, dynamically inferring context-specific weightings, and deploying autonomous agents for feature selection and continual learning. Furthermore, these studies do not address embedding-aware privacy tuning, generative data augmentation, or dataset-specific agent control, particularly in text-centric domains. In summary, while the current literature provides robust cryptographic and DP-based solutions for user privacy in smart grid and IoT environments, these approaches are largely static and lack scalability across domains. The integration of adaptive, context-aware agents capable of learning

from evolving user preferences and embedding nuances represents a critical direction for future research.

Within the electric mobility sector, the rapid proliferation of EVs, further accelerated by advancements in autonomous driving [150], has elevated the data privacy of autonomous EV owners as a central concern from a Responsible Artificial Intelligence perspective [151]. In related fields, recent studies such as [152] have extended metric DP to datasets representing three-dimensional rotations, thereby providing geometric guarantees for rotational datasets. While this work compares data privacy and utility tradeoffs across Laplace and Bingham mechanisms, it does not incorporate adaptive or agentic mechanisms for qualitative-to-quantitative translation, contextual inference, or continual learning, nor does it address the utility of semantic or textual data. The concept of Individual Differential Privacy (iDP) introduced by [156] offers personalized sensitivity-based guarantees that improve utility preservation, but it functions as a one-time optimization and lacks adaptive intelligence or the ability to learn from user semantics. The absence of agentic recalibration and semantic awareness restricts its applicability in dynamic privacy-utility scenarios. Additionally, ref. [153] presents Bayesian Differential Privacy (BDP) for correlated data, advancing theoretical leakage modeling. Nevertheless, BDP remains a static probabilistic framework without context-aware thresholding, agent-driven feature identification, or mechanisms for continuous feedback adaptation. The Utility-optimized Local Differential Privacy (ULDP) approach by [154] enhances mean estimation under local DP through utility-aware mechanisms. Although effective for numeric data, ULDP does not support dynamic parameter learning, qualitative user mapping, or context-specific adaptation, and it lacks integration of semantic or agentic reasoning. The study by [155] aims to improve user comprehension of privacy settings through interactive visualization. While notable for its human-centered design, this approach does not autonomously learn from user input or update its models over time, remaining a user-driven interface without adaptive feedback or semantic calibration. The most recent comparative analysis by [157] empirically evaluates DP and synthetic data mechanisms across multiple datasets. Although this study establishes a valuable

empirical baseline, it does not provide automation for mechanism selection, text-focused semantic evaluation, or continuous privacy-utility optimization through adaptive agents.

Outside of specific domains, Ref. [158] introduced the "Minimum Fragment Access" metric and a set-partitioning formulation, marking a significant theoretical and computational advance in database-level privacy management. However, this approach operates within a fixed optimization regime and does not incorporate learning or adaptive mechanisms. The model cannot learn from user-defined utility semantics or dynamically adjust its weighting. Similarly, [159] proposed a fuzzy-logic and differential privacy-based framework for privacy-preserving tabular data sharing. While methodologically robust, this framework is limited by its rule-based adaptation mechanism rather than a continuously learning agentic process. The fuzzy inference system relies on predefined membership rules rather than reinforcement or supervised learning informed by evolving user feedback. Although the framework models trust-level adaptation, it does not implement agent-driven feature selection or context-aware categorization to dynamically reweight data characteristics across domains. The absence of a generative synthesis component for data augmentation and the lack of autonomous parameter evolution based on semantic embedding characteristics further limit its ability to dynamically calibrate privacy–utility. Overall, the literature demonstrates considerable theoretical and empirical progress in privacy preservation, but remains largely static, mechanism-centered, and domain-specific. As summarized in Table 2.1, no existing work offers an autonomous agentic framework capable of translating qualitative user intent into quantitative thresholds, learning contextual weightings from real-world feedback, performing agent-driven feature discovery to identify true utility determinants, or enabling continuous semantic calibration between privacy and utility through embedding-aware adaptation.

Table 2.1: Literature review summary of existing works in optimizing data utility and privacy considering five key comparative metrics, where P denotes partially addressed while N denotes not addressed.

Referenced Literature	Adaptive Qualitative to Quantitative Translation	Context-Aware Optimization	Feature-Driven Utility Discovery	Dynamic Privacy-Utility Balancing	Adaptive Intelligence through Continuous Learning
[142]	N	N	N	N	N
[143]	N	N	N	N	N
[144]	N	P	N	N	N
[145]	N	N	N	N	N
[146]	N	P	N	N	N
[147]	N	P	N	N	P
[148]	N	P	N	N	N
[149]	N	N	N	N	N
[152]	N	P	N	N	N
[156]	N	P	N	N	N
[153]	N	P	N	N	N
[154]	N	P	N	N	N
[155]	N	P	N	N	N
[157]	N	N	N	N	N
[158]	N	P	N	N	N
[159]	N	P	P	N	N

2.7 Summary

The literature reviewed in this chapter indicates that the rapid adoption of EVs, driven by advancements in battery and charging technologies, supportive policies, economic incentives, and broader social and environmental pressures, is transforming transportation and creating both operational challenges and new opportunities for the power grid. Key EV-enabled value streams, including EV charging management, fleet optimization, V2G and V2H services, parking-lot energy management, and distribution-grid monitoring for EV loads, can enhance reliability, reduce costs, and increase flexibility when effectively coordinated. Furthermore, the growing reliance on data and AI, particularly in AVs, introduces essential requirements for RAI and a broad range of AI risk domains. Safety, security, and accountability emerge as recurring themes that must be proactively

managed to maintain trust and support adoption. The analysis of privacy-utility tradeoffs highlights the need for adaptive, context-aware methods that translate qualitative intent into quantitative controls, dynamically balance privacy with analytical value, and continuously learn from real-world feedback. These capabilities establish privacy-preserving intelligence as a foundational enabler for scalable EV and AI-based mobility ecosystems.

3 | Analysis and Modeling of Value Creation Opportunities and Governing Factors for Electric Vehicle Proliferation

The previous chapter reviewed existing studies that identified and, in some cases, examined factors influencing EV proliferation. For instance, while a number of studies considered vehicle purchase price to be of lower importance, reference [16] emphasized the significance of lower purchase tax. Regarding operational costs, references [14, 17] focused exclusively on EV efficiency, whereas [15, 16] identified vehicle maintenance and electricity price as dominant factors. Although each study adopted a distinct approach to analyzing these factors, none provided a comprehensive analysis of all variables affecting EV market share. For example, reference [13] identified charging time as the most critical factor for advancements in battery technology. The emergence of new factors, such as the impacts of COVID-19 and transactive energy initiatives, as well as the diminishing influence of others due to technological progress, policy changes, and jurisdiction-specific conditions, further complicates the analysis.

This chapter provides a more comprehensive approach than previous related work and considers the following factors collectively:

1. Number of EVSE (N_{EVSE}), which includes EV charging infrastructure and also studying it against the number of gas stations.
2. Improvement in battery technology (IBT), including lower charging time and battery-density.
3. Drop in the cost of raw material (CRM) for EVs to be considered over time.
4. EV driving range (EVDR).
5. Safety concerns related to EVs (SCEV).
6. EV charging infrastructure interoperability (CII) across different EV models.
7. Purchase cost (PC) for the vehicle, which includes purchase price, tax, related subsidies, and economic exemptions.
8. Operational cost (OC) for the vehicle, which includes electricity and gas prices and their increase over time, EV and ICV efficiencies, average vehicle life and yearly mileage.
9. EV-related privileges (EVP) such as dedicated lanes, HOV access, and others.
10. Exemptions for EV (EVE).
11. Available EV models (AEVM).
12. GHG emissions targets (ET) as directed by regulators to motivate automaker investments in EVs and vehicle buyers' purchase decisions towards an EV.
13. Short-distance trips (SDT) more common since the COVID-19 pandemic.
14. Refueling convenience (RC), which comes along with EV also charging at home versus ICV only at public gas stations.
15. EV usage in vehicle fleets to become mainstream (FM).

16. Transactive energy initiatives (TEI) increasingly leveraging EVs, bringing them to mainstream.

Table 3.1 summarizes the existing literature that identified different factors and, in some cases, provided the approach to studying EV proliferation through those factors.

Table 3.1: Related work which identified or analyzed EV proliferation factors.

Factors	**[12]**	**[13]**	**[14]**	**[15]**	**[16]**	**[17]**	**[18]**	**[19]**	**[20]**	**[21]**	**This work**
N_{EVSE}		✓	✓	✓	✓		✓		✓		✓
IBT		✓							✓		✓
CRM								✓	✓		✓
EVDR			✓		✓	✓		✓	✓		✓
SCEV					✓						✓
CII							✓				✓
PC	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
OC			✓	✓	✓	✓	✓	✓	✓	✓	✓
EVP	✓			✓			✓				✓
EVE				✓			✓				✓
AEVM	✓				✓						✓
ET	✓		✓				✓			✓	✓
SDT						✓					✓
RC											✓
FM							✓				✓
TEI										✓	✓

3.1 Dynamic Modeling of EV Adoption

It is important to highlight that based on the literature survey, none of the existing work thus far provides an approach to support the following objectives:

1. Identify a comprehensive approach to understanding all the factors which impact EV proliferation and consequently impact EV market share with respect to their traditional counterparts i.e., ICVs.

2. Study and categorize different value streams in the EV domain.

3. Provide an approach to analyze EV proliferation in the light of different EV value streams.

4. Deliver a flexible approach to analyze EV proliferation to aid in decision making for investment in technology advancements or informing policy changes for a regulatory regime.

This research provides a comprehensive approach to achieving the above-mentioned objectives. While Section 2.2 examined existing literature to identify the factors which impact EV proliferation, Section 2.3 presented a survey across academic studies and industry initiatives to generate the list of value streams. The current Section 3.1 and the following Section 3.2 help achieve the remainder of the objectives by providing a flexible approach to analyze EV proliferation under different scenarios, which could help in decision making around investments and policy making for EVs.

This section presents a model which helps study EV proliferation by leveraging all the contributing factors presented in Section 2.2. Each factor influences a vehicle buyer's decision to choose an EV versus an ICV, impacting the total market share of EVs at any given time. For simplicity, the proposed model assumes the availability of only two types of vehicles in the market, i.e., ICV and EV, and ignores other types, such as fuel-cell vehicles. Moreover, the modeling of each contributing factor reflects one possible approach where it could be further refined to more accurately represent individual cases. Figure 3.1 presents a high-level view of the model, and Figure 3.2 represents

its MATLAB implementation, highlighting EV proliferation factors, their collective impact on informing relative EV proliferation rate (REVPR), and consequently overall EV market share (EVMS) at any given point in time.

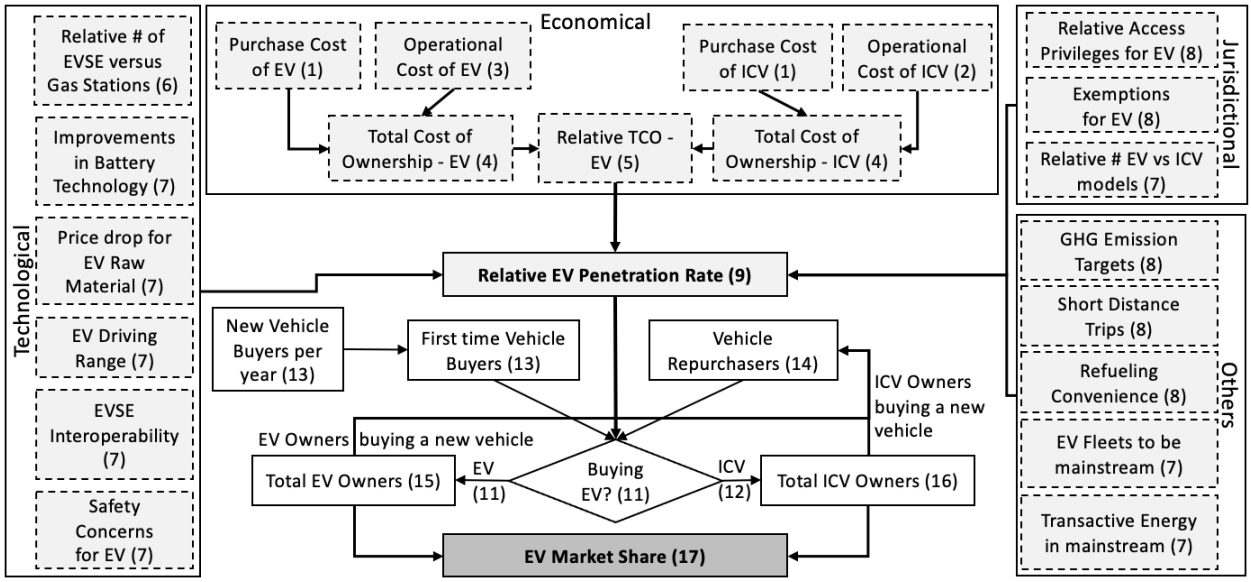


Figure 3.1: Logical model view depicting REVPR and EVMS calculations using EV proliferation factors.

3.1.1 Calculating REVPR Using EV Proliferation Factors

For economic factors, a vehicle's purchase cost (PC) is primarily a combination of the vehicle purchase price and taxes imposed at the time of purchase. Therefore, the purchase cost of a vehicle can be depicted as follows:

$$\mathrm{PC}_{\mathrm{Vehicle}} = \mathrm{Price}_{\mathrm{Vehicle}} + \mathrm{Tax}_{\mathrm{Vehicle}} \tag{3.1}$$

where Vehicle $\in$ {EV, ICV}.

Similarly, a vehicle's operational cost (OC) comprises mileage and maintenance costs. For an ICV, the mileage cost can be calculated using the cost of gasoline consumed and vehicle efficiency, whereas, for an EV, the electricity consumed towards charging the vehicle is considered along with

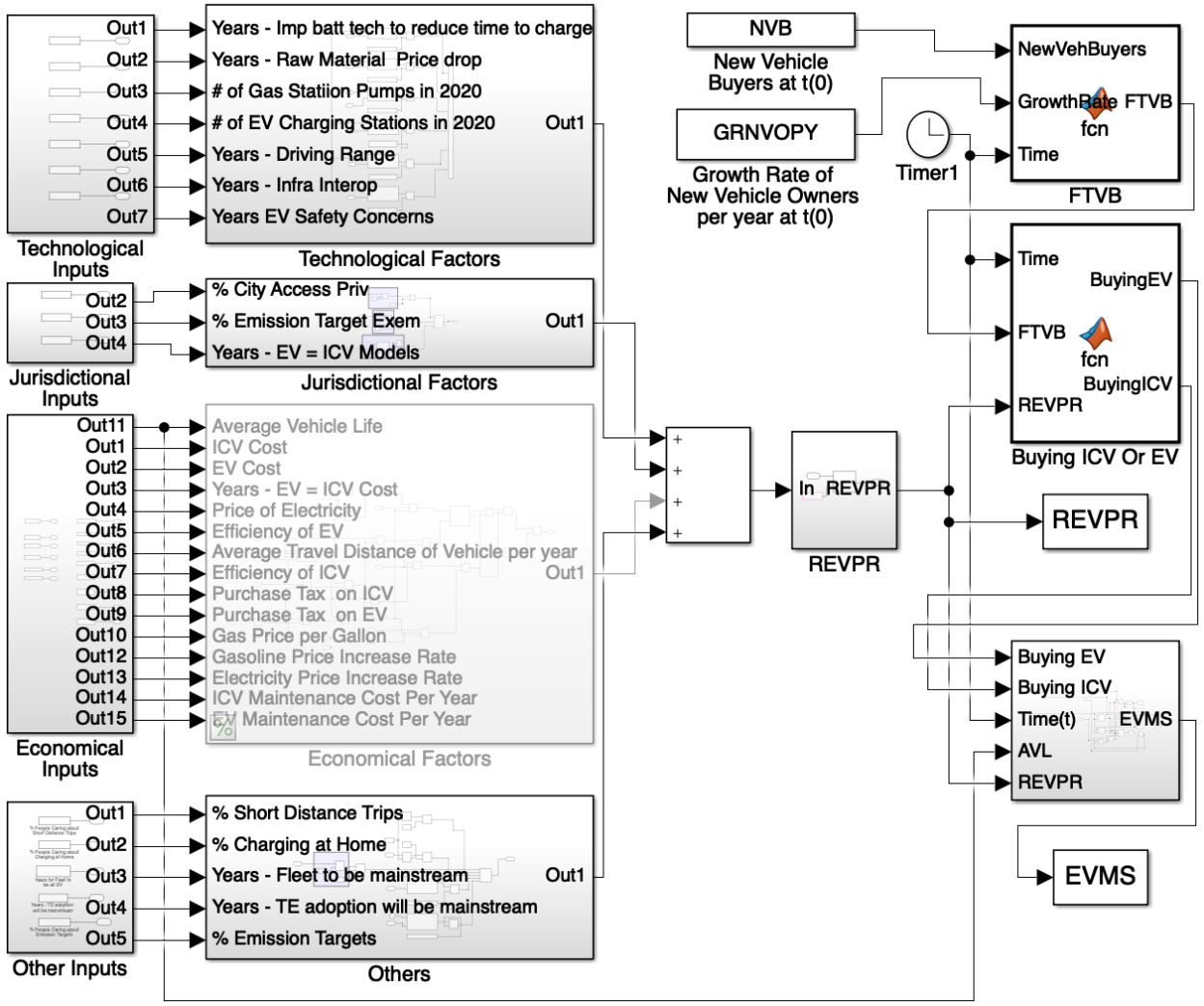


Figure 3.2: MATLAB-based dynamic model with configurable factors.

vehicle efficiency. The following set of considerations are made to simplify the model:

- One vehicle per vehicle owner or buyer;
- A fixed vehicle mileage per year (VMPY);
- Same average vehicle life (AVL) for both ICV and EV;
- Constant efficiencies for ICV (Eff_{ICV}) and EV (Eff_{EV});
- A fixed rate of increase for gasoline (RIGP) and electricity (RIEP) prices;

- A fixed maintenance cost per year for ICVs ($CMPY_{ICV}$) and EVs ($CMPY_{EV}$) respectively;
- Zero cost for accidental damages.

Therefore, considering AVL, VMPY, Gasoline Price (GP) per gallon, Electricity price (EP) per unit, RIGP, RIEP, Eff_{ICV}, and Eff_{EV}, the operational costs for an ICV and EV can be formulated as:

$$OC_{ICV} = AVL*VMPY*GP(t) * (1+(RIGP*t))*Eff_{ICV} \tag{3.2}$$

$$OC_{EV} = AVL*VMPY*EP(t) * (1+(RIEP*t))*Eff_{EV} \tag{3.3}$$

The total cost of ownership of a vehicle (TCO) over its lifetime can be calculated as

$$TCO_{Vehicle} = PC_{Vehicle} + OC_{Vehicle} \tag{3.4}$$

where the Vehicle could be either EV or ICV.

Therefore, the relative total cost of ownership for EV with respect to ICV ($RTCO_{EV}$) can be depicted as:

$$RTCO_{EV} = \frac{TOC_{EV}}{TOC_{ICV}} \tag{3.5}$$

Considering technological factors, the relative rate of EVSE growth compared to the number of gas stations can be represented by the following equation:

$$EVSE(t) = ((1 + R)^t * N_{EVSE}(0))/N_{GS}(0) \tag{3.6}$$

where R = rate of increase for EV Charging Stations, t = projection years, $N_{EVSE(0)}$ = number of EV charging stations at the beginning of the simulation, and $N_{GS(0)}$ = the number of gas stations at the beginning of the simulation. This equation can be used to study relative EV infrastructure for

any geography.

As discussed, EV Driving Range (EVDR) is expected to improve in the coming years and may become a negligible factor in decision-making when choosing between an EV and an ICV. This study introduces the term Saturation Time (ST), which denotes the period after which a factor no longer significantly influences the choice between an EV and an ICV. For example, the Saturation Time for Driving Range (ST_{DR}) can be modelled as a linear increase from 0 to 1 over a specified number of years within a particular region. Similarly, advancements in battery technology (IBT), which helps reduce the time required to fully charge an EV, improvements in EV charging infrastructure interoperability (CII), increased availability of EV models compared to ICVs (AEVM), the mainstream adoption of EV fleets (FM), and the growth in the number of Transactive Energy initiatives globally (TEI) are all anticipated to improve substantially in the near future. As a result, these factors are also expected to become insignificant in the decision-making process for purchasing an EV over an ICV. Furthermore, safety concerns related to EVs (SCEV), primarily due to battery technology, and the costs of raw materials (CRM) such as batteries and copper used in motors, are expected to decrease significantly and similarly become less influential in consumer decisions.

Therefore, all the above-mentioned factors influencing a vehicle buyer's decision over time will collectively steer them toward buying an EV. Hence without loss of generality, their effect can be modeled as a linear rise from 0 to 1 in their respective saturation times (ST) measured in the number of years for a particular geography and represented as follows:

$$\mathrm{X(t)} = \begin{cases} 1, & t > ST_{\mathrm{X}} \\ t/ST_{\mathrm{X}}, & 0 < t < ST_{\mathrm{X}} \\ 0, & t = 0 \end{cases} \tag{3.7}$$

where X ∈ {EVDR, IBT, CII, AEVM, FM, TEI, SCEV, CRM}

On the other hand, automobile manufacturers have been striving to meet GHG emission targets

(ET) as established by different jurisdictional mandates such as EU's Climate action plan [45], Zero-Emission Vehicle (ZEV) Program in California, USA [160], and China's New Energy Vehicle Credits regulation [161]. However, from an EV buyer's perspective, it only influences the decision making of those who are willing to or can afford to pay higher prices in favor of EVs, especially when there is a cheaper ICV counterpart available today. This subset of the population's size and percentage could also vary depending on the per-capita income across different geographies. On the other hand, EV-related privileges (EVP) and exemptions (EVE) discussed before are temporary measures that will only be available until EVs become mainstream and may not significantly influence decision making for a vehicle buyer after that period.

Due to COVID-19, people's preferences have changed recently, as highlighted by some recent studies [12, 162, 163]. Many people now perform short-distance trips (SDT) due to workplace closures and remote-working arrangements. Due to pandemic-related safety reasons, refueling convenience (RC) has also become a new priority where; the people who own EVs increasingly prefer to charge at home versus going to public charging stations. To some degree, the safety concerns also promote EV proliferation as ICV owners do not have that choice and mostly have to go to a public gas station to refill their vehicles. If the pandemic sustains, [164] suggests that these factors may increasingly influence the vehicle buyer's decision going forward. Irrespective of whether the pandemic factors sustain or not, their impact will only influence a subset of the overall vehicle buyers, and hence they can be mathematically represented as:

$$\mathrm{Y}(t) = \frac{\%\text{ of vehicle buyers caring about "Y" at a given point in time (t)}}{\text{Total number of vehicle buyers at a given point in time (t)}} \tag{3.8}$$

where Y $\in$ {ET, EVP, EVE, SDT, RC}

As highlighted in Figure 3.1, the above set of factors represented by Equations (3.5)–(3.8) together helps define the relative EV proliferation rate (REVPR) with respect to ICVs at any given point in time. REVPR is calculated as a weighted sum of these factors where each factor's weight

could be assigned on a scale of 0 to 1, representing 1 as high, 0.66 as a medium, 0.33 as low, and 0 as insignificant based on the geographical applicability of each factor. The weighted sum method is widely adopted and is applicable for scenarios where it is necessary to calculate a composite objective function that combines multiple objective functions into one scalar [165]. Therefore, the REVPR can be defined as the following:

$$\text{REVPR}(t) = \sum_{i=1}^{M} w_i f_i(x(t)) \tag{3.9}$$

where, $f_i(x(t))$ represents individual factors impacting EV proliferation and w_{i} represents their corresponding weight, satisfying

$$\sum_{i=1}^{M} w_i = 1, \quad w_i \in (0, 1) \tag{3.10}$$

3.1.2 Calculating EVMS Using REVPR

The REVPR defined using the above approach signifies a vehicle buyer's decision making in choosing an EV over an ICV, where a higher value of REVPR represents a higher probability of choosing an EV. REVPR influences the decision making of both first-time vehicle buyers (FTB) and vehicle repurchasers (VR). Therefore, at any given point in time (t), the number of EV buyers (EVB) and ICV buyers (ICVB) can be represented as:

$$\text{EVB}(t) = \text{REVPR}(t) * (\text{FTVB}(t) + \text{VR}(t)) \tag{3.11}$$

$$\text{ICVB}(t) = (1 - \text{REVPR}(t)) * (\text{FTVB}(t) + \text{VR}(t)) \tag{3.12}$$

where FTVB can be calculated using the growth rate of new vehicle buyers per year (GRNVPY) over an initial set of vehicle buyers:

$$\text{FTVB}(t) = \text{FTVB}(0) + \int_0^t \text{New Vehicle Buyers}(0) * (1 + \text{Growth Rate})^{\tau} d\tau \tag{3.13}$$

and VR can be considered as a combination of EV owners (EVO) and ICV owners (ICVO) whose vehicle's average lifespan (AVL) is completed and are going to buy a new vehicle:

$$\text{VR}(t) = (\text{EVO} + \text{ICVO}) \text{ where AVL} \geq 10 \text{ years} \tag{3.14}$$

Consequently, the total number of EVO and ICVO can be calculated using the following:

$$\text{EVO}(t) = \text{EVO}(0) + \int_0^t \text{EVB}(\tau) d\tau \tag{3.15}$$

Similarly, the total number of ICV owners (ICVO) can be calculated using the following:

$$\text{ICVO}(t) = \text{ICVO}(0) + \int_0^t \text{ICVB}(\tau) d\tau \tag{3.16}$$

Finally, the EV market share (EVMS) can be obtained using the following:

$$\text{EVMS}(t) = \frac{\text{EVO(t)}}{\text{EVO(t)} + \text{ICVO(t)}} \tag{3.17}$$

3.1.3 Quantifying Value Streams Using REVPR

At a high level, the concept of value streams is qualitative. However, as it helps in EV-related investments and policy-making decisions, it is essential to quantify value streams for efficient decision making and course correction where needed. One way to quantify value stream is by assessing it

against a "desired" EV proliferation that is required for a value stream to be considered "effectively" realized based on a policy or business decision. This research employs binary classification to quantify the "effectiveness" of the value stream, which further helps assess the efficacy of a previous decision as good or bad or make new decisions regarding policies and investments. Therefore, the relationship between a value stream and REVPR can be represented as follows:

$$\text{VS(t)} = \begin{cases} 1, & REVPR(t) \geq REVPR_{\text{desired}} \\ 0, & REVPR < REVPR_{\text{desired}} \end{cases} \tag{3.18}$$

where VS is a value stream such as EVCM, EVFMO, and others as described in Section 2.3, REVPR(t) is the relative EV proliferation rate at a given point in time t, and $REVPR_{\text{desired}}$ is assigned based on a policy or business decision, for a value stream to be considered "effective".

It is essential to note that applicable domain knowledge must be leveraged to identify the most influential factors shaping the specific value stream. For instance, the Electric Vehicle Fleet Management Operator (EVFMO) achieves greater effectiveness through favourable policy decisions and increased investments that support electric vehicle fleet expansion, which, in turn, leads to selecting a lower value for ST_{FM}. There is no universally optimal method for selecting an ST value; rather, it requires qualitative estimation informed by domain expertise and the extent of policy and investment efforts. A lower ST value necessitates immediate policy action and investment, whereas a higher ST value allows for a more relaxed approach. Mathematically, the REVPR Equation (3.9) is influenced by the X(t) Equation (3.7), which depends on ST. Thus, the relationship between the value stream and ST can be logically expressed as follows:

$$\text{VS(t)} : REVPR \longrightarrow ST_{\text{X}} \tag{3.19}$$

3.2 Model Validation and Simulation

The model presented in the previous section is simulated for multiple scenarios which analyze the relative EV proliferation and total market share across jurisdictions to study the timelines for each EV value stream. The base set of values for all factors, as highlighted in Table 3.2, are obtained from multiple sources [1, 9, 25, 26, 27, 42, 43, 166, 167, 168] where, without loss of generality, the geography is selected as USA and the base year for simulation as 2020. The model can be applied across any jurisdictional and regulatory boundary to study EV proliferation as long as the supported data is available. Going forward, as new factors emerge, the model will require updates to include those as part of the analysis.

3.2.1 Scenario 1: Comparative Analysis to Assess the Impact of Additional Factors on EV Proliferation and Market Share with Respect to the Previous Related Work

This scenario focuses on assessing the contribution of this work by considering additional EV proliferation factors not covered in previous studies, as highlighted in Table 3.1. To realize that, the base model factors are simulated against the scenario where the specific factors based on Table 3.1 were considered for respective studies and the rest were eliminated by setting their weight to zero. Figures 3.3 and 3.4 present the projections for REVPR and EVMS. It is evident that the REVPR and EVMS generated from the previous studies differ significantly among themselves, highlighting the importance of additional factors considered in this research and the importance of a more comprehensive approach. Based on the values for the base scenario, the time when the EV proliferation will surpass the total number of ICV in the specific geography, i.e., EV(#) > ICV(#), is around 20 years, whereas, when specific factors are ignored as per previous works of literature, it is more than 30 years. However, it should be noted that the importance of this analysis could

be understood by changing the values of different factors and weights to inform decision making around technology decisions, investments, and policy making. A separate scenario is presented later to analyze this aspect further.

Table 3.2: Base values for scenario analysis.

Factor	Value		Factor	Value
AVL (years)	10		GP(0) in 2020 (USD/gallon):	2.419
GRNVPY(0) in 2020	0.018		EP(0) in 2020 (USD/unit):	0.2
VMPY (miles)	15,000		RIGP (%)	0.0223
ICV Cost (USD) in 2020	16,000		RIEP (%)	0.018
EV Cost (USD) in 2020	35,000		Eff_{ICV}	0.08
Tax_{ICV} (%)	10		Eff_{EV}	0.195
Tax_{EV} (%)	10		$ST_{IBT,\ CRM,\ EVDR,\ SCEV}$ in years	10
$CMPY_{ICV}$ in USD	500		ST_{FM} in years	15
$CMPY_{EV}$ in USD	170		$ST_{CII,\ AEVM,\ TEI}$ in years	20
Gas Stations in 2020: $N_{GS}(0)$	590,000		$ST_{CAP,\ ETC,\ SDT,\ CAH}$ (%)	50
EVSE in 2020: $N_{EVSE}(0)$	78,500		ST_{ET} (%)	100
New Vehicle Buyers in 2020: NVB(0)	17,000,000		$w_{IBT,\ CRM,\ CHI,\ SCEV,\ ECO}$	1
ICV in 2020: TINICV(0)	287,300,000		$w_{EVDR,\ CII,\ AEVM}$	0.66
EV in 2020: TINEV(0)	1,800,000		$w_{CAP,\ ETC,\ SDT,\ CAH,\ FM,\ TEI,\ ET}$	0.33

3.2.2 Scenario 2: Analyzing the Impact of Technology Improvements on EV Proliferation

The focus of this scenario is to assess the importance of technological factors as recent EV technology improvements impact multiple factors in our research, including IBT, CRM, EVDR, CII, and SCEV. The technological improvement projections depend on jurisdictional policies and geopolitical forces and may vary once they change. Therefore, it is vital to understand the impact of technology improvement variations on EV market share over time. As highlighted in Figure 3.5, the

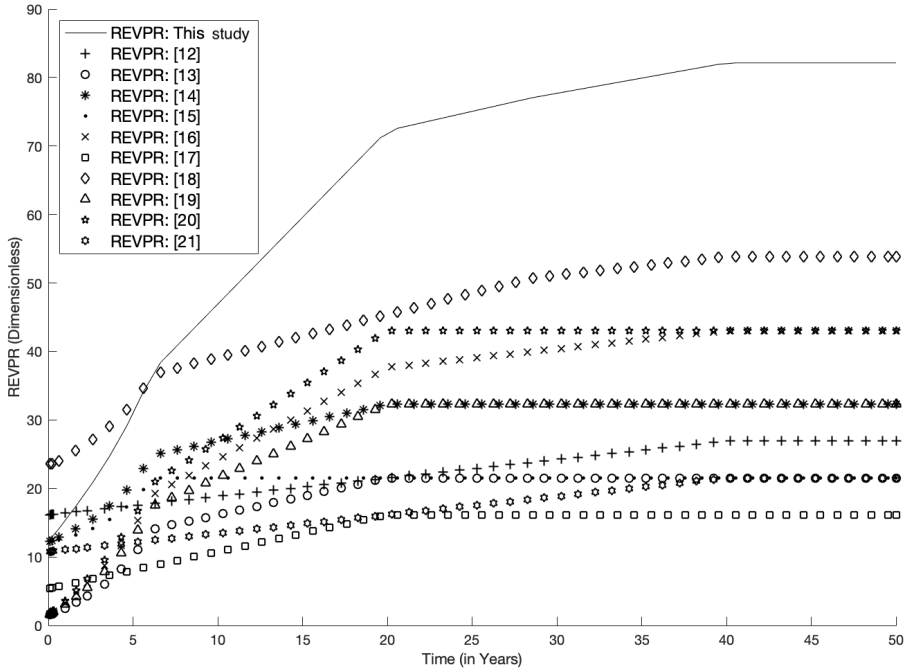


Figure 3.3: Relative EV Proliferation Rate.

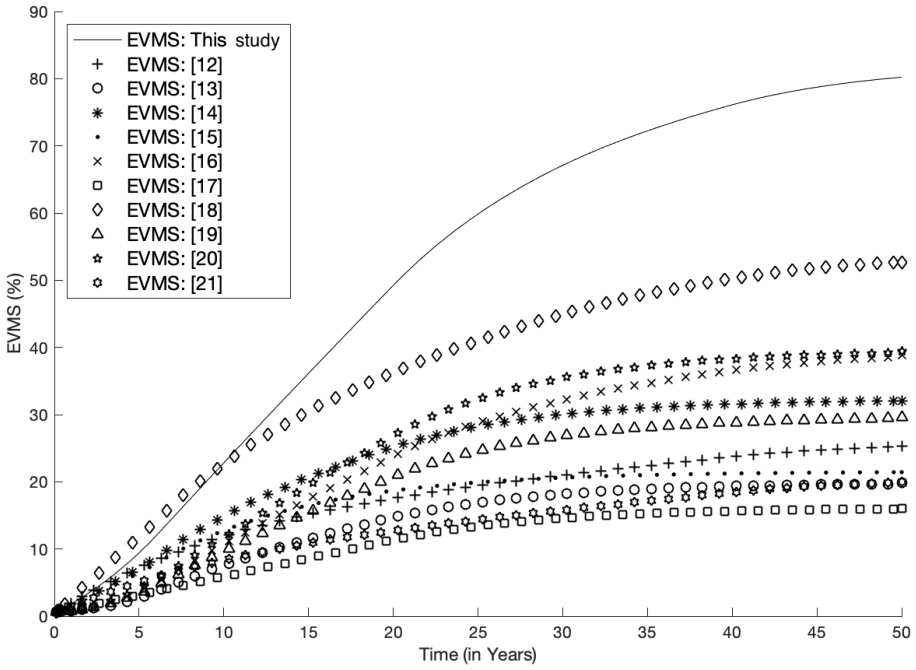


Figure 3.4: EV market share.

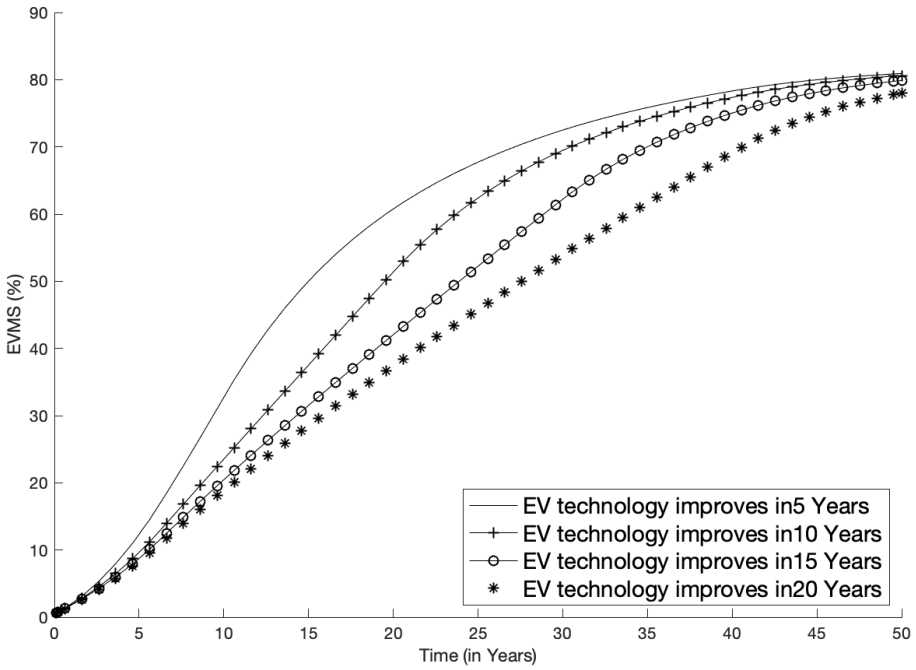


Figure 3.5: EVMS for different technology improvement trends.

quicker technology improvements, such as in less than five years, can help EVs capture the majority market share in just fifteen years from now.

3.2.3 Scenario 3: Approach to Assess the Impact of the COVID-19 Pandemic

A recent NREL report [9] studied the challenges due to the COVID-19 pandemic, including the resulting restrictions and economic downturn, and concluded that the EV charging industry did not get impacted as severely as other areas of the energy sector [169]. This scenario focuses on assessing this impact by evaluating the EVMS projections considering with and without pandemic factors. As represented by the EVMS values in Figure 3.6, the impact of the pandemic alone may not be noticeable soon, but its gradual impact over a more extended period will demonstrate a significant change in people's preferences towards buying EVs over ICVs.

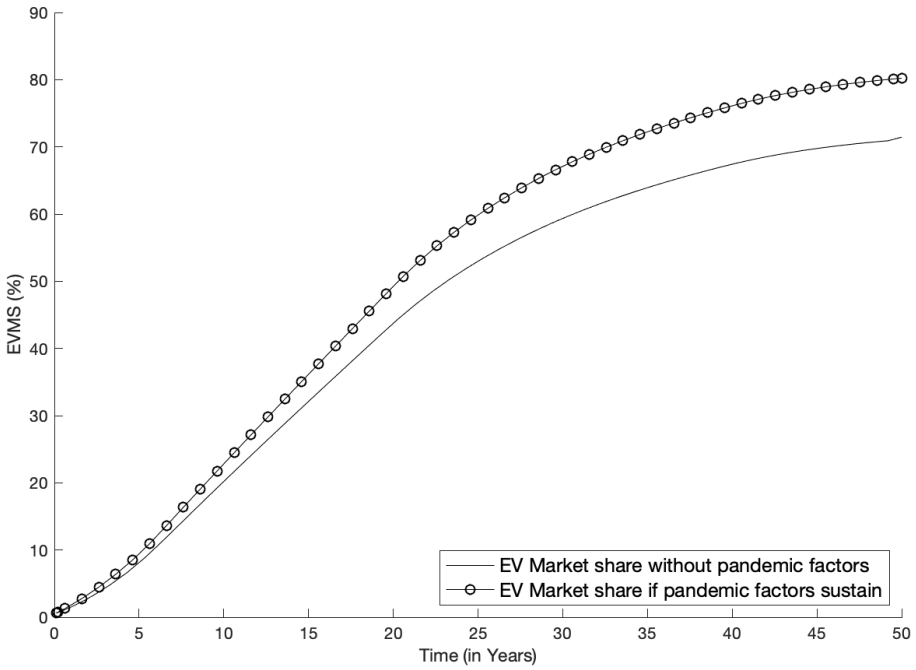


Figure 3.6: EVMS projections based on pandemic impacts.

3.2.4 Scenario 4: Demonstrating Framework's Applicability to Realize Value Streams Through EV Proliferation Factors

This scenario highlights the EV proliferation framework and model capabilities by demonstrating the interdependency between the EV proliferation factors and the value streams through simulations. A reference to the latest US administration goals of 50% EV sales by 2030 [170] is made to quantify value streams' "effectiveness" aspect and assign a realistic value to $REVPR_{desired}$. In other words, as REVPR represents EV sales over ICV in our model, setting $REVPR_{desired}$ = 50 will help us consider each value stream to be "effectively" realized, as described in Section 3.1.3. Moreover, leveraging the example from Section 3.1.3, a ST_{FM} = 10 will help ensure EV in fleets will have achieved its saturation point in 10 years (2020–2030) and will no longer act as a dominant factor in an EV buying decision. Therefore, assigning values of REVPR = 50 and ST_{FM} = 10 could be utilized to adjust the remainder of the values as provided against EVFMO in Table 3.3. A visual representation of this

example is provided in Figure 3.7.

Please note that since each EV proliferation factor impacts individual value streams differently, Table 3.3 represents one possible illustration of the same. As highlighted in Section 3.1.3, each ST is directly governed by decision making on technological investment, policy directives, and other aspects such as innovation breakthroughs. Therefore, it is evident that based on the domain knowledge, certain factors and corresponding values of ST can be made larger while others smaller on a case-by-case basis. For example, increased investment by EV manufacturers to expand the number of EV models will lower the ST_{AEVM}, directly impacting associated value streams and making them become "effective" sooner. In another case, a policy directive to subsidize FO may result in a sharp increase in the number of Fleet EVs, in turn reducing the ST_{FM}, thus rapidly generating a high value for the EVFMO value stream. In a third case, a sudden innovation breakthrough may help reduce the battery or copper prices, in turn lowering the ST_{CRM} and creating high-value across multiple value streams. Therefore, changing the policy or investment for a particular aspect will impact the corresponding ST, in turn helping achieve the desired value of REVPR in the desired time period.

The mapping in Table 3.3 represents an illustrative example that helps analyze how EV proliferation factors could work differently across each value stream to achieve the same goal of minimum 50% EV sales, as shown in Figure 3.7.

Table 3.3: Value-stream mapping with EV proliferation factors aid in analyzing the timelines to be met for each factor influencing individual value stream.

Value Stream	Tech Factors ($ST_{IBT, CRM, EVDR, CII, SCEV}$)	ST_{AEVM}	ST_{FM}	ST_{TEI}	Tax_{EV}
EVCM	10 years	20 years	15 years	10 years	10%
EVFMO	10 years	10 years	10 years	10 years	0%
V2(G/H/V)	5 years	20 years	15 years	10 years	0%
PEM	10 years except ST_{CII} = 5 years	20 years	15 years	10 years	10%
DGMEV	10 years	20 years	15 years	5 years	10%

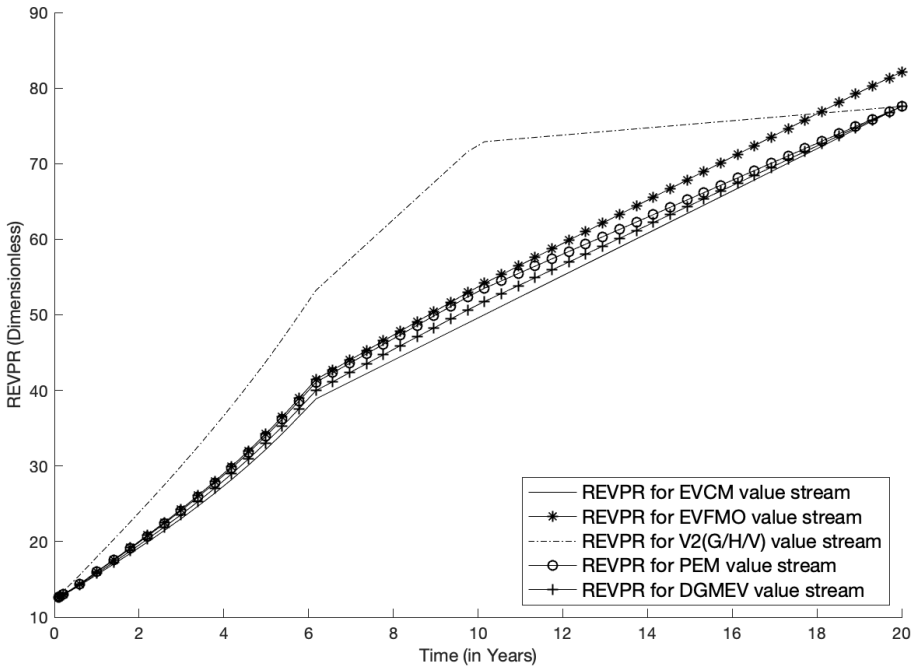


Figure 3.7: REVPR to achieve jurisdictional mandates.

3.2.5 Scenario 5: Comparative Analysis to Understand Model Applicability Across Different Countries and Jurisdictional Policy Regimes

This scenario highlights the model's applicability across different countries, where individual policies and locational decisions may impact EV proliferation differently. To keep the focus on understanding policy impacts on economic and other related factors, the following assumptions have been made to forecast the EV market share:

1. The selection of countries, such as Norway, Sweden, and the Netherlands, is based on the understanding that they have similar environmental sustainability focus, which drives policy directives, consequently impacting EV proliferation.

2. The USA's inclusion in this scenario study enables the comparison of a different jurisdictional regime partially driven by state policies and not just federal directives, as in the case of the

other three countries.

3. Values for the factors typically not impacted by government policies are kept the same across countries. Those include AVL, VMPY, CMPY_{ICV}, and CMPY_{EV}.

4. Based on the assumption that geo-political reasons such as energy (fuel and electricity) prices have a similar global impact, factor values such as RIGP (%) and RIEP (%) are kept the same across countries.

5. Assuming technology advancements propagated similarly across all countries, values for Eff_{ICV} and Eff_{EV} are kept the same across each country.

6. To ensure fair comparison across all the countries, parameters as $\text{ST}_{\text{IBT, CRM, EVDR, SCEV}}$, ST_{FM}, $\text{ST}_{\text{CII, AEVM, TEI}}$, $\text{ST}_{\text{CAP, ETC, SDT, CAH, ET}}$, and $w_{\text{IBT, CRM, CHI, SCEV, ECO}}$, w_{EVDR}, w_{CII}, w_{AEVM}, w_{CAP}, $w_{\text{ETC, SDT, CAH, FM, TEI, ET}}$ are also assigned equal values across countries.

Considering the above assumptions, the rest of the values highlighted in Table 3.4 are obtained from different sources [171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189]. These factors have a direct impact due to a policy directive, investment decisions, or other policy-related aspects and are, therefore, critical in producing the results in the scenario analysis.

The values in Table 3.4 are leveraged to analyze the policy impacts and individual decisions made across different countries. As depicted in Figure 3.8, Norway is already leading the pack in terms of EV market share due to its increased focus on policies such as a low tax on EVs as opposed to Sweden and Netherlands. Other factors, such as relatively higher gas prices and much lower electricity prices than its neighbors, also support high EV proliferation in Norway. The USA, on the other hand, is the smallest in EVMS% compared to the other three countries, given its vast network of gas stations and ICV numbers, despite being one of the largest uptakers of EVs globally.

Table 3.4: Values across different countries.

Factor	USA	Norway	Sweden	Netherlands
AVL (years)	10	10	10	10
GRNVPY(0) in 2021	0.02	0.024	0.018	0.1399
VMPY (miles)	15,000	15,000	15,000	15,000
ICV Cost (EUR) in 2021	15,095	40,000	35,000	33,000
EV Cost (EUR) in 2021	33,000	39,000	57,000	52,000
Tax_{ICV} (%)	10	30	21	25
Tax_{EV} (%)	10	5	22	16
$CMPY_{ICV}$ in EUR	400	400	400	400
$CMPY_{EV}$ in EUR	140	140	140	140
Gas Stations in 2021: $N_{GS}(0)$	620,000	1600	2900	4100
EVSE in 2021: $N_{EVSE}(0)$	113,600	19,300	16,335	99,500
New Vehicle Buyers in 2021: NVB(0)	14,900,000	176,276	301,006	320,000
ICV in 2021: TINICV(0)	287,300,000	5,354,451	171,573	224,000
EV in 2021: TINEV(0)	667,731	647,000	129,433	96,000
GP(0) in 2021 (EUR/gallon):	2.28	7.23	6.367	6.783
EP(0) in 2021 (EUR/unit):	0.19	0.13	0.25	0.32
RIGP (%)	0.0223	0.0223	0.0223	0.0223
RIEP (%)	0.018	0.018	0.018	0.018
Eff_{ICV}	0.08	0.08	0.08	0.08
Eff_{EV}	0.195	0.195	0.195	0.195
$ST_{IBT,\ CRM,\ EVDR,\ SCEV}$ in years	10	10	10	10
ST_{FM} in years	15	15	15	15
$ST_{CII,\ AEVM,\ TEI}$ in years	20	20	20	20
$ST_{CAP,\ ETC,\ SDT,\ CAH,\ ET}$ (%)	100	100	100	100
$w_{IBT,\ CRM,\ CHI,\ SCEV,\ ECO}$	1	1	1	1
$w_{EVDR,\ CII,\ AEVM,\ CAP}$	1	1	1	1
$w_{ETC,\ SDT,\ CAH,\ FM,\ TEI,\ ET}$	1	1	1	1

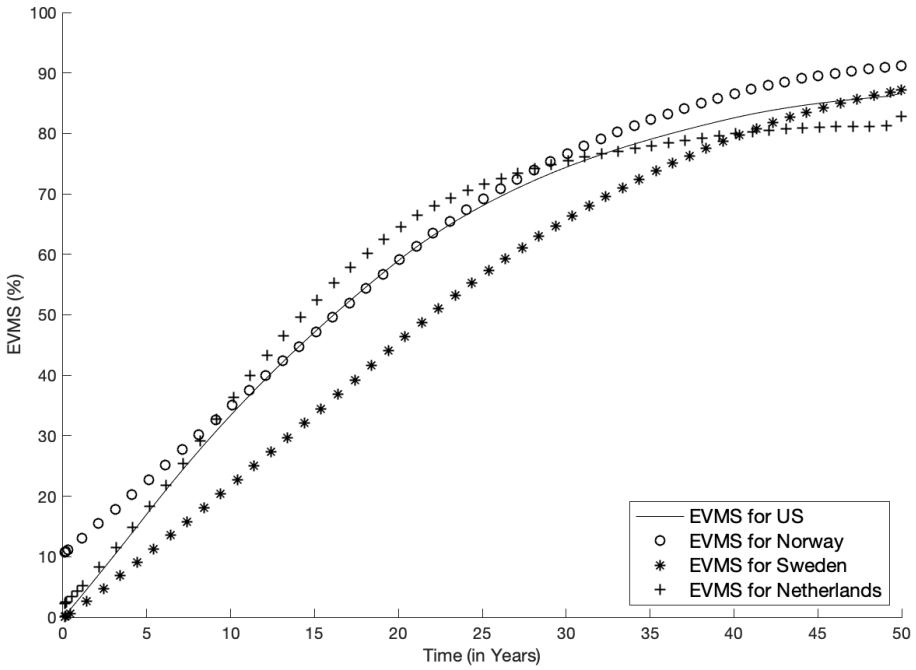


Figure 3.8: EVMS projections across different countries subject to different policy directives.

3.3 Key Considerations for Framework Applicability

It is important to highlight that while the scenario analysis through model validation and simulations demonstrates the framework's flexibility, the results presented in the analysis were entirely based on the model equations and input data. Therefore, to apply the model accurately and obtain meaningful results, careful attention must be paid to the following:

- Identification of additional factors arising from evolving geopolitical perspectives, technological advancements, and policy-driven directives;
- Evaluation of each factor and its relative significance within the specific scenario;
- Verification of the accuracy of input data for independent variables;
- Adjustment of model coefficients across various equations; and

- Where applicable, evaluation and adjustment of model equations when alternative relationships are observed between dependent and independent variables.

3.4 Summary

This chapter presented a configurable dynamic modeling framework that connects a comprehensive set of EV proliferation factors to a relative EV proliferation rate and, subsequently, to time-varying EV market share. This approach enables both explanatory analysis and forward-looking projections. By explicitly incorporating economic, infrastructure, technological, and policy drivers, along with their interactions, within a unified set of equations, the model addresses limitations in previous research that considered only narrow subsets of influences and lacked generalizability across changing conditions. The framework is validated through multiple simulation scenarios that include the effects of previously overlooked factors, technological advancements, pandemic-like disruptions, value-stream realization, and cross-country policy comparisons. These scenarios demonstrate how stakeholders can utilize the same core structure to investigate hypothetical situations and quantify potential outcomes. The findings highlight the need to continually update factor selection, weights, coefficients, and input data to ensure the model remains accurate and relevant for decision-making as real-world contexts evolve.

4 | Responsible AI Framework for AI-based AVs: Addressing Bias and Fairness Risks

Addressing all the AI risks, as previously discussed in Section 2.5, requires a holistic approach involving collaboration between key stakeholders, including policymakers, industry leaders, researchers, and the public, who will be end-users of the AVs. Involvement from each stakeholder is essential to develop robust governance frameworks, ethical guidelines, and technical solutions that prioritize safety, security, privacy, fairness, and societal well-being in the deployment of AVs. Therefore, this thesis proposes a system-level AI Risk Identification and Mitigation approach for an AI-based AV system by implementing a holistic RAI framework, as depicted in Fig. 4.1. The first layer of the conceptual framework brings together the nine crucial AI risk domains discussed earlier, necessitating comprehensive risk identification and mitigation strategies. Each risk domain comprises multiple risks that impact the real world usage of AI-based AVs, which need to be addressed in the context of AV's functionality. The proliferation of cloud and edge computing, combined with the usage of automation and AI further improved by Generative AI technology, poses additional risks that were not present in traditional non-AI-based vehicles.

The objective of the second layer is to identify different risks across the entire AI lifecycle for each risk domain under consideration. This layer in the framework presents three stages of

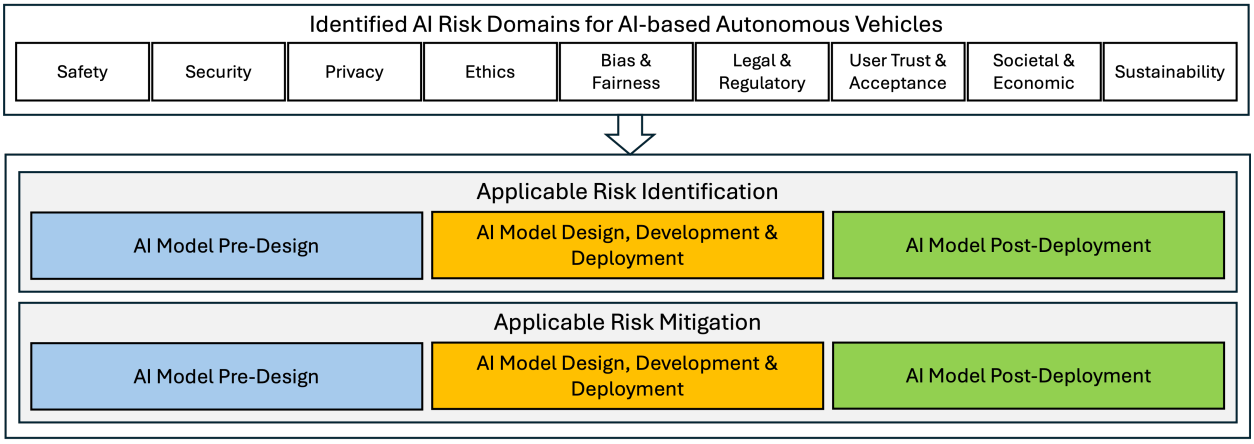


Figure 4.1: RAI Framework for AI-based AVs

RAI intervention for the AI lifecycle: the pre-design stage for the AI model; AI model design, development, and deployment; and post-AI model deployment. These stages play pivotal roles in ensuring the applicable risks are identified and mitigated at each stage of the AI development lifecycle, bringing efficacy, reliability, and safety to AV systems. The AI risks in the pre-model design stage include risks during data collection and processing activities comprising of data acquisition, storage, and pre-processing of data necessary for training, testing, and operation of the AI system. This stage also includes risks creeping in due to poor system design practices, wherein the overall architecture and functionality of the AV system are conceptualized and defined. AI model design, development, and deployment activities constitute the second stage, focusing on creating and optimizing AI algorithms and models tailored for AV tasks. The third stage includes all the activities following the deployment of the AI model, where the AI model is maintained and regularly tuned for optimal performance. This stage involves evaluating and validating AI model performance and effectiveness in real world scenarios achieved through ongoing management, maintenance, and monitoring of the AI system to ensure continued functionality and performance.

The third layer presents the components for effective risk mitigation across each stage discussed in the second layer. Please note that these mitigations are not a one-time effort; in most cases, they are continuous improvements for each functionality of an AV that leverages AI. For example,

at the AI system and data level, mitigation efforts could involve implementing and continuously improving cybersecurity measures, data encryption techniques, and access controls to safeguard against security breaches and privacy violations. Similarly, strategies, including incorporating fairness-aware algorithms, bias detection mechanisms, and diverse training data to mitigate biases and ensure equitable outcomes, need to be continuously worked upon at the AI model design and development level. Finally, at the AI model monitoring level, continuous monitoring, auditing, and evaluation of AI systems are essential to detect and address potential risks, anomalies, or deviations from expected behavior, enhancing AI-based AV systems' reliability, transparency, and safety.

The proposed framework could help achieve the objectives of an RAI-based AV system spanning across each phase of AI design, development, deployment, and real world AV operations leveraging AI. It is essential to highlight that the framework's second and third layers apply to all AI risk domains highlighted in Layer 1. However, it is crucial to consider the different risks under each domain to elaborate on the framework for all risk domains. For the simplicity of the framework's application and with the objective of not losing generality, this study selected bias and fairness as the risk domain to identify and mitigate each applicable bias risk applicable to AV's AI lifecycle. The application of the remaining risk domains is considered a future research opportunity.

4.1 Describing applicable Bias and Fairness Risks for AVs

Bias and fairness are concepts shaped by societal norms and are contingent on the context, reflecting individuals' or organizations' values, ethical standards, and legal requirements. To ensure that AI systems serve the best interests of both people and society, it is essential to establish a clear definition of fairness within a specific context and to be capable of evaluating whether a system operates impartially. Furthermore, AI development teams must effectively communicate this definition and ensure the adherence of AI systems to it. Equitable treatment of all groups by AI models is paramount, and proactive measures should be taken to mitigate potential biases and

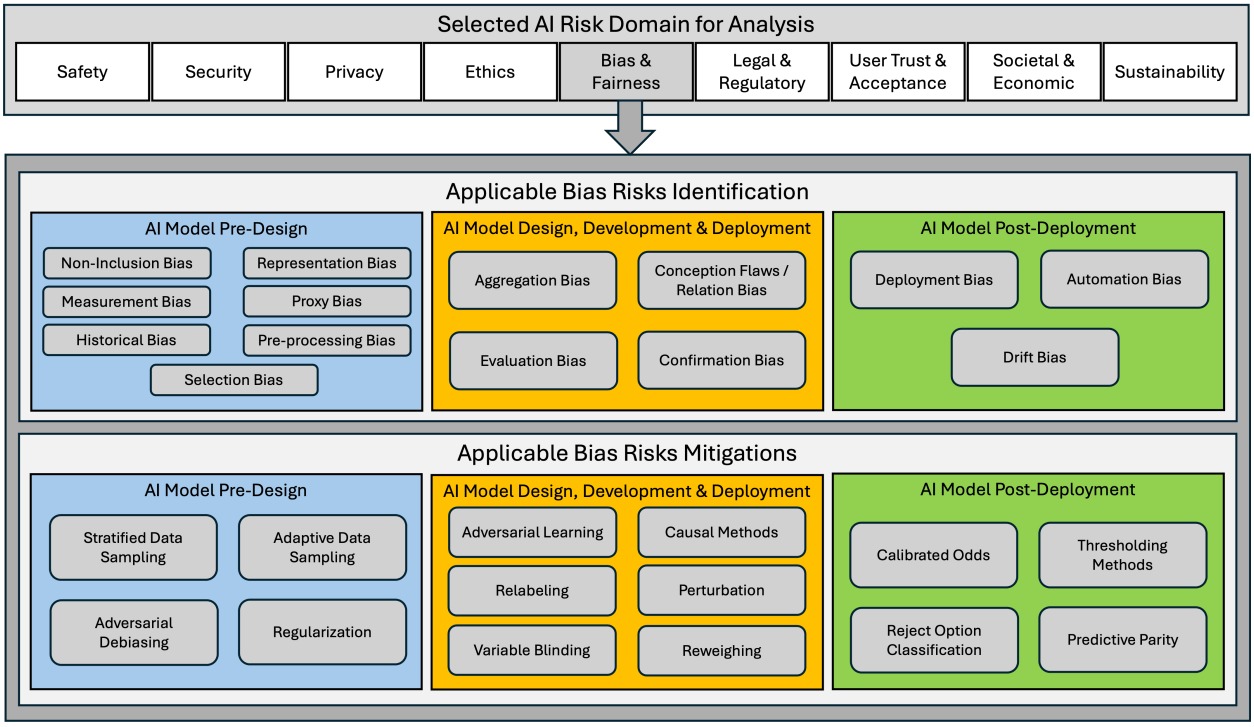


Figure 4.2: Bias Risks and Mitigations across the AI lifecycle by leveraging the RAI Framework for AI-based AVs

prevent adverse outcomes, such as discrimination based on gender, race, or ethnicity. Fig. 4.2 presents important bias risks and mitigations useful across different stages of the AI lifecycle for an AI-based AV which are described in the following sections.

4.1.1 Stage 1: Pre-design stage for AI Model

During Stage 1, the bias risks could creep in as early as during the AI system design conceptualization due to non-inclusion bias or during the data collection through representation bias. It is essential to address these biases early by incorporating diverse stakeholders (prevent non-inclusion), ensuring diverse, representative datasets (avoid representation and selection bias), using transparent metrics and minimizing proxies (prevent proxy and measurement bias), and acknowledging historical inequalities and compensating for them during data processing (mitigate historical and pre-processing bias). Table. 4.1 highlights individual risks during stage 1 and their analysis through

Table 4.1: Bias and fairness risks during pre-design stage for AI-based AV

Risk	Description	Identification	Mitigation Direction
Non-inclusion bias [190]	This bias risk creeps in when early design choices unintentionally exclude certain user groups, not due to data but due to limited considerations in system conception. For instance, an AV system that only supports specific languages may inadvertently restrict access for non-speakers, creating barriers to equitable use.	One method to detect non-inclusion bias is to explicitly define the target audience and consider groups that may be unintentionally excluded. This helps ensure inclusive design from the start. Post-deployment, monitor usage and gather feedback to identify and address any accessibility issues specific groups face.	To address non-inclusion bias, apply inclusive design principles by prioritizing accessibility and supporting multiple languages. Regularly monitor and collect feedback to make necessary adjustments, particularly for underserved groups. Engage a diverse design team early to anticipate and resolve usability issues.
Representation bias [191, 192, 193]	This risk arises when a training dataset lacks diversity, failing to reflect the broader population the ML model serves. This limited data leads to bias, causing the ML model to favor certain patterns and reducing its ability to generalize effectively.	Two main methods to detect representation bias are (1) checking for equal representation of subgroups within a feature and (2) ensuring subgroup distributions match those in the broader population. Choosing the best approach requires consulting domain experts for the specific application.	While representation bias does not invalidate data or models, it should be recognized because it can cause performance disparities across subgroups. To address this, researchers can collect more diverse datasets or use methods such as oversampling or undersampling to ensure balanced subgroup representation.
Measurement bias [192, 193, 194]	This risk refers to systematic bias arising in data collection, recording, or analysis, causing a deviation from the true value of the variable. It can occur at any stage, from instrument design to data interpretation.	To detect measurement bias, review measurement procedures to ensure they accurately capture needed information. Comparing sample results with true values from a larger population or reference standard can highlight discrepancies. Additionally, comparing outcomes from different methods or instruments measuring the same variable can reveal potential biases.	To address measurement bias, use standardized, validated tools appropriate for the context, verify results with multiple measurements, and apply blinded or double-blinded procedures when observer effects may occur, such as the Hawthorne effect. Disclose any limitations or potential sources of bias when interpreting data and drawing conclusions.
Proxy bias [192]	In data science, a proxy variable is indirectly related to the variable of interest. Proxy bias occurs when this proxy is used as a substitute for an unmeasured variable in a statistical or ML model, potentially leading to inaccurate conclusions.	A method to detect proxy bias involves scrutinizing the variables for making predictions and assessing whether they genuinely pertain to the desired outcome or are associated with any other input variable.	To address proxy bias, prioritize variables that directly correlate with the desired outcome. If only proxy variables are available, consider collecting additional data or using feature engineering to create variables more closely linked to the outcome.
Historical bias [193, 195]	This risk arises from existing biases in the world that are reflected in the data. As the world changes, models trained on outdated data may become obsolete.	To detect historical bias, review available data to determine if it differs from the intended outcome or is prone to change.	To address historical bias, augment data for underrepresented scenarios or adjust the model's output by assigning appropriate weights.
Pre-processing Bias [196]	Pre-processing bias occurs when data cleaning, transformation, or preparation introduces distortions. Altering specific data or features during these steps can result in biased outcomes.	Pre-processing bias can be identified by reviewing decisions made during data cleaning, transformation, and preparation. Assess whether data or features are omitted or altered in ways that may introduce bias, and consider the impact on specific groups or variables.	To address pre-processing bias, conduct sensitivity analyses using different pre-processing techniques and evaluate their effects on results. Data scientists should apply business knowledge to ensure data is synthesized and refined appropriately.
Selection Bias [197]	This bias occurs when samples are not randomized, leading to non-representative data. Key types include sampling bias, where data does not reflect the broader population; reporting bias, which misrepresents event frequencies; participation or non-response bias, resulting from skewed responses due to non-participation; and survivorship bias, which considers only selected individuals and excludes those who did not qualify.	To detect selection bias, confirm that the dataset accurately represents the intended population. Review the data's origin, its alignment with the target population, and any relevant subgroups or distributions. Assess missing features and their effects on groupings. Identifying representation bias, which may result from selection bias, is also helpful.	If the sampling process does not cover the entire population, mitigate selection bias by re-sampling with a different method if possible, seeking additional datasets to represent the remaining population, and revising the study framework and conclusions as needed.

identification and mitigation strategies. Furthermore, these risks can also compound and amplify each other, leading to dangerous oversight in decision-making, suboptimal performance in real world conditions, and disproportionate risks for underrepresented groups. These biases work together to make the AI system unreliable and inequitable, eroding public trust and increasing the liability for AV manufacturers. The following is an analysis of the cumulative impact of multiple biases, emphasizing how they interact and create compounding risks.

1) Non-Inclusion Bias & Representation Bias: If non-inclusion bias prevents diverse voices (e.g., rural or disabled individuals) from being considered during the design phase, the resulting datasets may also have representation bias. These biases together can cause the AVs to perform poorly in underrepresented conditions, such as narrow roads in rural areas or interactions with pedestrians using wheelchairs, making the system unsafe or non-functional for certain groups.

2) Measurement Bias & Proxy Bias: These two biases can interact when the AV system relies on proxy data that amplifies existing inaccuracies. For example, in cases when lower-income neighborhoods are used as a proxy for high-risk driving conditions, measurement bias could cause sensors to misidentify obstacles in these areas (e.g., darker-toned objects at night). In this scenario, the AI-based AV may behave more cautiously or erratically in such locations, leading to discriminatory or unsafe behavior and reducing trust and adoption.

3) Historical Bias & Pre-Processing Bias: If historical data already contains systemic biases (e.g., more accidents recorded in neighborhoods with poor infrastructure), and pre-processing techniques focus only on cleaning or smoothing typical patterns, important edge cases might be lost. This results in models reinforcing existing inequalities, with AVs less prepared for rare but critical scenarios—like sudden potholes or jaywalking in underdeveloped areas.

4) Selection Bias & Representation Bias: These biases compound each other, as a non-representative data selection limits the scope of what the AV can accurately predict and react to. For instance, if the training data disproportionately comes from sunny highways, the AV will struggle with scenarios like snowy rural roads. This creates an unsafe system that performs well in ideal environments but

fails when faced with real world complexities.

5) Proxy Bias & Non-Inclusion Bias: A lack of diverse stakeholder input (non-inclusion bias) combined with reliance on proxies can lead AVs to make flawed assumptions about safety and risk. For example, when proxies like income levels or infrastructure quality are leveraged to predict safe driving conditions, the AV may neglect to account for the unique needs of cyclists or individuals crossing in less-regulated areas. This could result in dangerous oversight, disproportionately harming those already marginalized.

6) Measurement Bias & Historical Bias: If AV sensors struggle to capture certain objects accurately (measurement bias), and these objects are also underrepresented in the training data (historical bias), the problem is magnified. For example, when past data has fewer records of bicycles or small electric scooters, and the sensors used by the AV are not optimized to detect them. In that case, the AV will perform poorly in environments where such vehicles are common and could result in safety risks or even fatal accidents.

4.1.2 Stage 2: AI-model design, development, and deployment

This second set of bias risks, include aggregation bias, conception flaws, evaluation bias, and confirmation bias, which could appear due to insufficient technical considerations during the AI model design, development, and deployment activities. These biases, as described in Table. 4.2, could be identified in different ways and require specific mitigations before the model is deployed to an AV. Additionally, similar to Stage 1, the interaction of these biases during the AI model design and development stage could create compounding risks that can have serious real world consequences. These biases affect how data is aggregated, how models are conceptualized and evaluated, and how feedback is interpreted, leading to: overconfidence in AV capabilities, resulting in unsafe deployments; inability to generalize across diverse environments, causing unpredictable behavior; lack of testing coverage, where critical edge cases are neglected; and loss of trust and reputational damage following deployment failures. The cumulative impact of these biases interacting with each

Table 4.2: Bias and fairness risks during AI-model design, development, and deployment stage for AI-based AVs

Risk	Description	Identification	Mitigation Direction
Aggregation bias [192, 193, 198]	It occurs when data aggregation or summarization leads to information loss and potentially skewed outcomes. This process can obscure variations in more detail, resulting in biased conclusions. Relying solely on aggregated data can lead to erroneous deductions. Additionally, model aggregation bias arises when a universal model fails to consider distinct groups or examples, assuming a consistent mapping from inputs to labels across all data subsets, which may not reflect reality.	To identify aggregation bias, it is crucial to assess whether data is being overly summarized, potentially masking important subgroup differences, and involves examining how data is aggregated and ensuring that critical variations, such as demographic, environmental, or contextual differences, are not lost in the process. Collecting disaggregated data and performing sensitivity analyses can help uncover hidden biases early, ensuring the model accounts for diverse real world scenarios, which is especially vital for AVs operating across varied environments and populations.	To mitigate aggregation bias, selecting an appropriate level of granularity that aligns with the research goals is crucial. Conducting sensitivity analyses at different aggregation levels and examining disaggregated data can help address this issue. Additionally, maintaining transparency about the chosen aggregation level and any limitations or biases introduced during the aggregation process is essential when interpreting and presenting findings.
Conception flaws [195]	Conception flaws, or relation bias, refer to biases from how a research inquiry, hypothesis, or study design is conceptualized. These biases occur when the research question is poorly framed or incomplete, leading to potentially biased conclusions or interpretations of the findings.	Identifying conception flaws in AI involves thoroughly reviewing AI systems' design, development, and implementation. This includes assessing foundational assumptions, concepts, and data to uncover biases or constraints. Engaging diverse stakeholders throughout development, testing, and validation provides valuable perspectives and helps detect potential flaws.	To address conception flaws in AI, proactively tackling biases and erroneous assumptions during design and development is crucial. Engaging diverse stakeholders throughout these stages helps mitigate potential flaws.
Evaluation bias [192, 193]	This bias pertains to the existence of biased or flawed assessment techniques or criteria employed to gauge the performance or efficacy of AI systems. It can result in inaccurate or insufficient evaluations of an AI system's performance, as the assessment methods may not fully reflect the system's real world capabilities.	Detecting evaluation bias in AI requires thoroughly assessing the evaluation techniques, criteria, and datasets used to gauge an AI system's effectiveness. This includes ensuring the methods are comprehensive, impartial, and relevant to real world settings. Comparing the distribution of training and testing datasets can confirm their representativeness and check for drift bias. Additionally, examining potential biases in testing datasets helps identify evaluation bias.	To address evaluation bias, it is essential to utilize varied and inclusive test datasets that faithfully represent the real world environment where the system will be deployed. Additionally, continually updating and improving evaluation techniques and criteria based on feedback and real world performance can help minimize evaluation bias.
Confirmation bias [199]	Confirmation bias denotes a flawed decision-making process that reinforces preexisting notions or beliefs without a comprehensive and impartial assessment of the evidence. This bias can result in validating information or findings that align with existing beliefs while dismissing or disregarding contradictory evidence.	Confirmation bias can be detected by carefully scrutinizing the decision-making procedures, data selection, interpretation methods, and validation techniques employed throughout the design and development phases.	To counteract confirmation bias, it's essential to use diverse datasets and robust validation techniques, foster openness to varied viewpoints, anonymize data, and assemble a diverse team of experts and stakeholders from the start. Additionally, scrutinizing funding sources helps ensure impartial evaluation.

other can lead to systemic failures, unintended behaviors, and safety risks. The following list analyzes how the interaction between aggregation bias, conception flaws, evaluation bias, and confirmation bias affects the AV development lifecycle. These compounding risks can be mitigated broadly by diverse, context-sensitive data collection to avoid aggregation bias, scenario planning with diverse stakeholders to challenge assumptions and prevent conception flaws, comprehensive evaluation frameworks that account for edge cases and atypical conditions, and employing independent review processes to identify and mitigate confirmation bias.

1) Aggregation Bias & Conception Flaws: If developers design the AV model under flawed assumptions (conception flaws) and combine diverse datasets without proper contextual weighting (aggregation bias), the system will perform poorly in specific scenarios. For example, assuming that braking patterns for urban and rural settings are identical and merging their data without accounting for differences can result in unsafe braking behavior on rural roads. Such interactions make the AV brittle, with significant performance gaps that are hidden during development but surface dangerously during deployment.

2) Evaluation Bias & Confirmation Bias: Evaluation and confirmation biases can reinforce each other, creating a feedback loop where the AV flaws go unnoticed or are dismissed. For example, if developers expect their AV model to handle urban traffic well, they may design tests that primarily cover urban scenarios. Even if the model performs poorly in unexpected rural environments, the focus will remain on validating its urban performance. As a result, the vehicle could be prematurely deployed in settings for which it is unprepared, compromising safety and trust.

3) Aggregation Bias & Evaluation Bias: Aggregating data from different sources (e.g., multiple countries) without proper adjustments may give the appearance of a robust model, but if evaluations emphasize typical or average scenarios, edge cases will be neglected. For example, if the evaluation criteria focus on highway driving under good weather conditions, the AV may perform well in tests but struggle in snow-covered urban environments. This interaction results in overconfidence in AI model capabilities and can lead to deployment failures in untested conditions.

4) Conception Flaws & Confirmation Bias: When conception flaws guide design and confirmation bias leads developers to reinforce these flaws, the AV model can fail catastrophically. For instance, if developers assume that all pedestrians will wait at crosswalks, they may ignore testing scenarios where pedestrians jaywalk. Even if the model encounters such cases during testing, confirmation bias can cause developers to treat them as edge cases and dismiss them. The consequence is an AV system prone to unpredictable failures in real world environments, where jaywalking or unexpected pedestrian behavior is common.

5) Aggregation Bias & Conception Flaws & Evaluation Bias: This three-way interaction can produce a model that appears well-tested but fails to handle variability in real world scenarios. For example, the AV may behave erratically or dangerously in countries with informal traffic norms, such as drivers ignoring lane markings or pedestrians crossing at random points. This case could lead to unsafe deployments, particularly in underrepresented environments, and expose the AI model to legal and reputational risks.

6) Confirmation Bias & Aggregation Bias: This interaction can result in false confidence in the AV's performance. For example, developers may aggregate driving data from both clear and rainy conditions without giving proper weight to the rainy scenarios. During the evaluation, confirmation bias may cause testers to focus on successful cases under clear weather, ignoring poor performance in the rain. This scenario leads to an over-optimistic assessment, and the AV could fail catastrophically in adverse weather, compromising public safety and damaging trust.

4.1.3 Stage 3: Post-AI Model deployment

Biases could penetrate post-deployment of AI models in an AV, i.e., during AI model operations, and could be identified through appropriate monitoring as highlighted in Table. 4.3. As the AV encounters new challenges, performance can degrade over time, especially if there is inappropriate reliance on automation. However, the cumulative impact of post-deployment biases can lead to serious safety risks, operational inefficiencies, and loss of trust, all of which must be proactively

Table 4.3: Bias and Fairness risks post AI-model deployment stage for AI-based AV

Risk	Description	Identification	Mitigation Direction
Deployment bias [193, 200]	Deployment bias in AI systems occurs when a model's performance differs in the real world compared to its development or testing phases. This bias can arise from discrepancies between training data and the deployment environment, changes in user behavior, or unexpected interactions with other systems. Deployment bias may lead to gaps between expected and actual performance, resulting in unintended consequences or inaccuracies in decision-making.	Comprehensive monitoring and evaluation are essential to identify deployment bias in AI systems. This includes comparing the model's performance in real world scenarios to its performance during the development and testing phases, as discrepancies may indicate deployment bias. Analyzing feedback from users and stakeholders can also offer insights into any unexpected behaviors or performance issues that occur after deployment.	Mitigating deployment bias involves proactively adapting the AI model for real world conditions. Continuous monitoring and retraining with updated data from the deployment environment help the model adjust to changing patterns over time. Techniques like domain adaptation, which fine-tunes the model with deployment-specific data, can enhance performance. Regular audits, performance reviews, and transparency measures are also essential for effectively identifying and addressing deployment bias.
Automation bias [129]	Automation bias is the tendency of human operators to over-rely on automated decision-making processes, even when these processes are flawed. This bias occurs when individuals trust AI algorithms without critically assessing their accuracy, potentially leading to decision-making errors. As a result, operators may ignore contradictory information or neglect to intervene, mistakenly assuming the automated system is infallible.	Identifying automation bias in AI systems requires close monitoring of human-AI interactions and decision-making processes. This includes observing how users engage with the AI, their reliance on automated recommendations, and whether they override or disregard AI outputs. Additionally, collecting user feedback on their perceptions of the AI system's reliability and trustworthiness can offer valuable insights into the presence of automation bias.	Mitigating automation bias requires combining technological solutions with human-centered strategies. Key strategies include providing transparent information about the AI system's limitations, implementing decision-support tools for verifying AI-generated recommendations, and fostering a culture of skepticism and critical thinking through ongoing training on AI capabilities and limitations. These measures can reduce reliance on automation and promote balanced decision-making.
Drift bias [130, 201]	Drift bias refers to the phenomenon where the performance of a trained model deteriorates over time due to changes in the data distribution or underlying environment. This bias can occur when the assumptions made during model training no longer hold true in the deployment environment, leading to inaccuracies or prediction errors.	Identifying drift bias in AI systems involves continuously monitoring and analyzing model performance over time. This can be done by comparing historical and recent performance metrics to spot discrepancies. Concept drift detection algorithms can also help identify shifts in data distribution, indicating drift bias. Additionally, gathering feedback from users and stakeholders can provide insights into environmental or behavioral changes affecting model performance.	To mitigate drift bias, proactive measures are essential for adapting the AI model to changing conditions. This includes periodically retraining the model with updated data to maintain its relevance and accuracy. Techniques like online learning, which incrementally update the model as new data arrives, can help it adapt to evolving patterns. Additionally, integrating feedback loops and monitoring mechanisms into the system architecture allows for continuous evaluation and refinement to reduce the impact of drift bias over time.

managed through careful monitoring, diverse testing, regular model updates, and human oversight. The following list explains each bias and explores how their cumulative impact can lead to dangerous or unintended outcomes once the AV is in real world use.

1) Deployment Bias & Automation Bias: When an AV is deployed in conditions different from its training environment (deployment bias), it may encounter unfamiliar situations where it performs poorly (e.g., unmarked rural roads, unusual pedestrian behavior). If automation bias leads human operators or passengers to rely too heavily on the system's judgment without intervening when needed, the consequences can be severe. For example, in bad weather or unpaved roads, the AV might make incorrect decisions, such as failing to detect obstacles or misjudging turns. Because users assume the AV knows what it is doing, they may fail to take corrective actions, resulting in collisions or unsafe situations.

2) Drift Bias & Automation Bias: As environmental conditions evolve—such as new construction, seasonal changes, or sensor wear—the AV's original model may become outdated, leading to errors or delays in decision-making (drift bias). Meanwhile, if users develop automation bias, they may stop monitoring the vehicle's behavior and ignore warning signs (e.g., the AV struggling to handle new road layouts or missing traffic signs). Small performance degradations can accumulate over time without manual intervention, leading to near-misses or even critical failures. This is especially risky in edge cases for which the AV was not initially trained, such as sudden lane closures or newly introduced driving rules.

3) Deployment Bias & Drift Bias: When an AV is deployed in new environments (deployment bias), the system may already struggle with unfamiliar scenarios—such as different pedestrian behaviors in certain regions. Over time, if the environment further evolves (e.g., road infrastructure changes, updated speed limits, or sensor degradation), drift bias can magnify these challenges. For instance, a vehicle trained for urban settings with clear road signs may perform poorly in rural environments where markings are faint or missing. As conditions gradually drift, the gap between the model's expected environment and the real world widens, leading to frequent errors or unsafe behavior. In

such cases, compounding deployment and drift biases can result in the AV making inconsistent or unpredictable decisions, which can be highly dangerous—such as failing to yield at unexpected crosswalks or misinterpreting temporary road barriers.

4) Automation Bias & Deployment Bias & Drift Bias: In scenarios where an AV system designed and successfully tested for urban deployment is deployed more widely in suburban and rural areas (deployment bias), over time, environmental changes (e.g., new intersections, updated traffic signals, sensor wear) could lead to drift bias. However, due to automation bias, human operators may trust the AV system implicitly, assuming it can handle these evolving situations without manual oversight. This cumulative impact could potentially lead to, but not limited to, the following scenarios:

- Increased risk of accidents: As the AV encounters new or unexpected situations—such as pedestrians crossing outside crosswalks or weather-affected sensor readings—the model's performance deteriorates, but users fail to notice or intervene.
- Delayed corrective measures: Automation bias can also mean operators are slow to respond when the vehicle malfunctions, assuming the AV will correct itself. This case is hazardous during critical edge cases (e.g., sudden detours or emergency braking).
- Compounding performance degradation: Deployment in diverse, evolving environments combined with drift in system accuracy can cause progressive failures over time, eroding trust in the AV. For example, what starts as a minor misjudgment—like improper lane changes—can grow into systematic safety violations, such as incorrectly merging into traffic.

This three-way interaction creates a perfect storm of risk through deployment bias, which ensures the AV faces new challenges it was not trained for; drift bias, which indicates that even the trained behaviors gradually degrade; and automation bias, which causes users to over-rely on the AV, leading to missed opportunities for manual intervention when the system behaves incorrectly.

Furthermore, in AVs, edge computing is critical in processing and responding to real-time data from sensors, cameras, and other onboard devices. However, edge computing can introduce specific

Table 4.4: Bias and Fairness risks due to distributed data processing and edge computing

Risk	Description	Identification	Mitigation Direction
Latency-Induced bias [202]	Latency issues in edge computing can introduce delays in decision-making, mainly when data processing is distributed across multiple edge nodes with different response times. In AVs, where real-time response is crucial, latency differences across nodes can lead to differential outcomes in decisions, depending on which node processed the data and when.	In delayed data aggregation, bias occurs when faster-arriving sensor data (e.g., LiDAR over cameras in low light) disproportionately influences decisions, sidelining slower inputs. This can result in inconsistent response timing, such as earlier or more confident actions based on quicker data processing. Identifying this bias involves monitoring processing times across edge nodes and examining time-synchronization logs to detect discrepancies in sensor data arrival or processing.	Implementing edge-node level synchronization buffers to manage data flow ensures that inputs from various sensors are time-aligned before processing. This could help minimize response timing discrepancies, reducing bias due to latency variations. Similarly, implementing latency-aware model calibration that integrates real-time latency metrics into model adjustments to balance decision timing across nodes could also help mitigate this bias.
Resource-Allocation bias [202]	Resource allocation bias occurs when specific edge nodes have more computational resources (e.g., higher processing power, memory) than others, leading to unbalanced processing capacity across nodes. In AVs, edge nodes closer to higher-powered servers may process data faster or more precisely than nodes with limited resources.	Performance metrics comparison of essential metrics such as accuracy and speed of model predictions across edge nodes with varying computational capacities and resource utilization monitoring and analysis to detect unequal processing capabilities across nodes could help detect this bias.	Implementing load-balancing algorithms to redistribute tasks to underutilized or overloaded nodes dynamically, resource standardization to equip nodes with comparable processing capabilities where feasible, and efficient compression techniques to use lightweight models or data compression to alleviate strain on underpowered nodes could help mitigate this bias.
Edge-Node Prioritization bias [202]	This bias occurs when specific edge nodes are assigned greater importance, either due to geographic proximity or hierarchical configuration within the edge network. If data from certain locations or sensor types is processed with higher priority, this setup can lead to biases in decision-making.	This bias can be identified by examining whether certain nodes consistently influence decisions disproportionately and performing sensor data contribution analysis to identify which edge nodes are prioritized during data aggregation and processing.	Some techniques to mitigate this bias include implementing federated learning across edge nodes to ensure fair contributions from diverse geographic and sensor data, introducing multi-note redundancy for critical data streams to prevent over-reliance on prioritized nodes, and balancing nodes' contributions by assigning weights based on underrepresented contexts or inputs.

biases due to the distributed nature of data processing. Factors like latency differences, resource allocation disparities, and edge-node prioritization can lead to biases in decision-making, affecting AV safety and fairness. Table. 4.4 specifically highlights latency-induced, resource allocation, and edge-node prioritization biases in this regard.

4.1.4 Compounding bias risks across Stages 1, 2, and 3

The cumulative impact of biases across the AI lifecycle of AVs, spanning pre-design, development, deployment, and post-deployment, creates compounding risks that grow over time. Flawed assumptions early in the design phase (non-inclusion and conception flaws) feed into biased models (aggregation and evaluation biases), leading to inappropriate deployments and over-reliance on automation. These risks become more severe post-deployment, as drift bias erodes performance over time, often going undetected due to automation bias. Addressing these interacting biases through proactive planning, diverse data collection, robust evaluation, and continuous monitoring is essential to ensure AV systems' safety, reliability, and equity in real world environments. The following are the cumulative impacts of biases as they interact across the three phases of the AI lifecycle for AVs.
1) During Stage 1, Non-inclusion bias at this phase creates gaps in data used later in model development, limiting the system's ability to handle edge cases (e.g., pedestrians using wheelchairs or unregulated intersections in rural areas). Further, conception flaws at the design phase result in biased model architectures or data-gathering strategies, leading to downstream aggregation bias (e.g., treating rural and urban driving data as interchangeable). Baked into the model from the start, these flaws can cause systemic bias that worsens during real world deployment. Finally, poor problem framing (conception flaws) at this stage propagates automation bias post-deployment by overestimating the reliability of AV systems in complex environments.
2) During Stage 2, interaction with pre-design flaws, aggregation bias reinforces the non-inclusion bias from the pre-design phase, as critical stakeholder needs like rural drivers or cyclists are absent from both data collection and model design. Further, overconfidence in deployment and evaluation

bias amplify the consequences of conception flaws and non-inclusion bias by introducing serious limitations in the model's safety assessments, giving developers a false sense of readiness to deploy. This also interacts with automation bias—deployment teams assume the model works well across all conditions, further underestimating risks. Finally, inappropriate deployment, when biases are not identified and addressed during development, deployment bias emerges, leading to AVs operating in environments they were not adequately trained for (e.g., deploying models trained in urban areas to rural regions). This contributes to post-deployment failures.

3) At Stage 3, accumulated design to deployment failures, aggregation bias from the design phase, and non-inclusion bias from the pre-design phase worsen drift bias post-deployment. If critical environmental or regional differences are not captured during development, the system's performance deteriorates quickly in these untrained conditions. Further, automation reliance in risky scenarios and automation bias post-deployment compounds these issues—users rely too much on the AV system, assuming it will handle unexpected situations that were not accounted for during training (due to selection and evaluation biases). Finally, delayed detection of drift, such as when operators trust the AV system too much, leads to drift bias, where the system's effectiveness decreases over time—and remains undetected, allowing small issues to grow into critical failures, such as misinterpreting new road signs or sensor degradation affecting obstacle detection.

4.1.5 Advanced Bias Detection Techniques

Identifying biases for AI-based AV systems is an ongoing process as new risks emerge due to the changing AV landscape and external factors like infrastructure, environment, etc. Therefore, advanced statistical methods—such as Bayesian inference, causal inference, Shapley values, HMMs, Gaussian processes, counterfactual fairness, and anomaly detection—enable comprehensive bias detection and mitigation across the AI lifecycle of AVs. These techniques provide deep insights into uncertainties, causal relationships, and hidden biases, ensuring AV systems are robust, fair, and adaptive throughout their development and deployment. Table. 4.5 analyzes advanced statistical

Table 4.5: Analysis and applicability of advanced statistical methods to identify biases in AI-based AV

Technique	Description	Applicability
Fair Representation Learning [203]	This technique helps assess and adjust for model bias, ensuring fairness across multiple groups by creating latent representations of input features that remove correlations with protected or sensitive attributes (e.g., race, gender).	This technique helps ensure pedestrian detection models work equally well across different demographic groups by learning bias-free representations of people in various environments.
Adversarial Debiasing [203]	This technique employs adversarial networks to detect biases by training the model to remove discriminatory patterns. The system learns to make accurate predictions while neutralizing biases linked to sensitive attributes.	This technique helps detect and correct biases in object classification, ensuring that the model treats pedestrians or cyclists uniformly regardless of environmental features (e.g., neighborhood wealth levels).
Shapley Values [204]	Shapley values, drawn from cooperative game theory, provide insights into feature importance by measuring the marginal contribution of each input feature to a model's predictions. It helps identify when certain features (e.g., ZIP code, lighting conditions) have an undue influence on predictions, potentially indicating proxy bias or selection bias.	This technique is useful for AV's pedestrian detection model as Shapley values could reveal that predictions could be disproportionately influenced by factors like weather conditions or regional differences, highlighting where bias may reside.
Hidden Markov Models(HMMs) and Temporal Modeling [205]	HMMs are effective in modeling sequential data, such as time-series sensor data collected from cameras, LiDAR, and radar in AVs. It helps identify patterns of change over time, such as sensor degradation or performance decay in specific environmental conditions.	HMMs could be used to track changes in object detection accuracy over time, detecting drift bias when the AV's performance decreases on rainy or snowy days due to evolving environmental factors.
Counterfactual Fairness and Perturbation Analysis [206]	Counterfactual fairness focuses on determining whether a model's prediction would have been the same if a sensitive feature (e.g., demographic factor) were changed. It helps identify causal bias by testing whether outcomes depend on variables like ethnicity, location, or traffic type.	This technique can help an AV system's decision to stop at a crosswalk, which can be tested using counterfactual analysis to ensure that pedestrian recognition works equally well across different demographic regions or lighting conditions.
Gaussian Processes (GPs) [207]	GPs are non-parametric models that quantify uncertainty in predictions, making them well-suited for environments where data is sparse or noisy. GPs allow the AV system to flag high-uncertainty scenarios, helping identify potential areas of evaluation bias (e.g., untrained driving conditions).	A GP-based AV model might detect that it performs poorly at night or on certain road types and alert developers to collect more data from those environments, reducing deployment bias.
Multi-Task Learning (MTL) [208]	MTL models learn to solve multiple related tasks simultaneously, leveraging shared representations to improve generalization. MTL can identify when bias from one task leaks into another, such as when vehicle detection performance influences pedestrian detection in biased ways.	An AV system trained for both object detection and collision prediction could use MTL to ensure that improvements in one task (e.g., detecting trucks) don't negatively impact another (e.g., recognizing cyclists).
Variance-Based Feature Attribution (VFA) [209]	This method measures the variance in predictions attributable to each input feature, helping detect features that introduce biases. This method employs bias detection via feature analysis, where if certain variables (like specific road types or weather conditions) dominate the prediction variance, selection bias or proxy bias is suggested in the model.	VFAs could be leveraged to help reveal scenarios like selection bias where urban traffic data for model training overly influences AV braking decisions.
Deep Kernel Learning (DKL) [210]	DKL combines deep neural networks with kernel methods to capture non-linear relationships and uncertainties. DKL can detect hidden biases in deep learning models used for object recognition by providing uncertainty-aware predictions.	When used in AV perception systems, DKL can identify when the system struggles with rare edge cases, such as nighttime pedestrians, reducing automation bias by alerting operators.
Anomaly Detection Techniques (e.g., Isolation Forests, Autoencoders) [211]	These methods detect outliers in data or predictions, which can reveal the presence of bias when the system performs poorly in unexpected scenarios. Bias detection via outlier identification or anomaly detection can highlight underrepresented scenarios in the training dataset (e.g., rare weather conditions) that could lead to biased predictions.	This is useful in scenarios where an isolation forest could detect that the AV system's performance drops significantly in snowy conditions, pointing to evaluation bias from inadequate testing during development.
Federated Learning for Regional Bias Detection [212]	Federated learning allows models to be trained on decentralized data (e.g., from different regions) without sharing raw data, making it ideal for identifying geographic biases. While detecting bias across regions, federated learning can identify when local models perform differently across regions, highlighting potential biases in the AV's general model.	Regional models trained through federated learning might show that the AV struggles with roundabouts in one city but not in others, revealing location-specific biases.

techniques that allow for dynamic modeling, fairness assessments, and detecting hidden biases at various stages, ensuring the system's robustness, equity, and safety across AI-based AV design, development, deployment, and post-deployment phases. By proactively leveraging these methods, developers can build more reliable, equitable, and safer AVs that perform well in a wide range of real world conditions.

4.2 Risk Mitigation Techniques for AI-based AV Systems

This section focuses on describing AI risk mitigations and their applicability across different stages of AI lifecycle as highlighted in Fig. 4.2. System and data-level mitigations within Stage 1 are typically addressed upfront. The techniques for Stage 2, i.e., during the model design, development, and deployment level, should be typically applied next, although they can sometimes be independently addressed. Finally, the mitigation for Stage 3 will always be applicable post-deployment of the AI model, which is always recommended to be applied after Stages 1 and 2 due to the reasons discussed earlier. The following sections describe techniques at each of the three levels.

4.2.1 Pre-Model Design - Data level mitigations

During pre-model design level, **Data Sampling** techniques play a crucial role in mitigating bias by ensuring that the training dataset is representative of the population. Along with those, Table. 4.6 highlights additional data level mitigations vital for a holistic perspective of risk mitigation before the model design and development commences.

Table 4.6: Pre-model design - data level bias risk mitigations for AI-based AVs

Technique	Description	Applicability
Stratified sampling [213]	This technique divides the population into distinct strata based on certain characteristics (e.g., age, gender, region) and then samples proportionately from each stratum. Stratified sampling improves the precision and reliability of estimates by ensuring representation from all subgroups, particularly for rare or minority groups.	This technique is widely used in opinion polls, market research, and epidemiological studies. However, stratified sampling requires prior knowledge of the population's stratification variables and may be computationally intensive for large datasets with numerous strata.
Adaptive sampling [214]	This is an evolving approach that dynamically adjusts the sampling strategy based on real-time feedback or evolving data characteristics. Unlike discussed stratified sampling, adaptive sampling iteratively selects data points based on the information gained from previous samples, optimizing the sampling process over time.	This technique is well-suited for streaming data, where the underlying distribution may change over time, or in scenarios where the cost of sampling varies across data points. Adaptive sampling has applications in anomaly detection, online learning, and real-time monitoring systems.
Adversarial debiasing [215]	Adversarial debiasing is a technique to reduce biases in AI systems, promoting fairness in decision-making. It trains a model to perform its primary task while minimizing bias by introducing an adversarial network during training. The adversarial network generates counterfactual examples to expose biases, while the primary network learns to predict both original and counterfactual examples, improving resistance to bias.	This approach is particularly applicable for AVs in scenarios where equitable decision-making (e.g., recognizing diverse pedestrian groups) and fairness across varied environments are critical for safety and inclusivity.
Regularization [216]	Regularization techniques in AI-based AVs mitigate bias by introducing penalty terms that discourage discriminatory behavior, focusing on fairness rather than hypothesis complexity. Strategies include minimizing empirical risk under fairness constraints, adjusting true and false positive rate (FPR) for protected groups, ensuring fairness stability, and incorporating counterfactual terms.	For AVs, regularization can improve equitable decision-making, such as accurate detection across diverse demographic and environmental conditions. However, caution is needed, as these methods may impact the model's robustness and generalizability in dynamic real world settings.

4.2.2 AI-model design, development, and deployment stage bias risk mitigations for AI-based AVs

Ensuring fairness in AI systems requires a comprehensive approach spanning model design, development, and deployment to mitigate bias effectively. During the design phase, fairness-aware objectives should be integrated by selecting diverse and representative datasets while identifying potential biases in data collection. To ensure equiDP model learning, the development stage should incorporate bias mitigation techniques such as reweighting, adversarial debiasing, and causal inference. DP. 4.7 highlights multiple mitigation techniques useful in mitigating risks during this stage.

Table 4.7: AI model design and development level bias risk mitigations for AI-based AVs

Technique	Description	Applicability
Adversarial learning [215]	Adversarial learning methods can mitigate biases in ML models as pre-processing or in-processing techniques, particularly concerning sensitive variables. In pre-processing, adversarial interventions reduce dependence on specific variables by modeling causal properties before and after intervention. This process, similar to an optical illusion for machines, introduces intentionally misleading inputs, such as images designed to confuse neural networks in classification tasks. Adversarial approaches are advantageous because they accommodate multiple fairness constraints and treat the model as a closed box, making them widely applicable. However, they can be unstable and challenging to train reliably, especially in transfer learning scenarios. Generative Adversarial Networks (GANs) with fairness considerations offer a promising solution by handling unstructured data and generating "unbiased" datasets, increasing their utility across diverse domains.	Enhances model robustness by generating adversarial examples that expose hidden biases in AV perception and decision-making. While this improves resistance to biased inputs, it can lead to overfitting and requires substantial computational resources, making real-time implementation challenging.
Causal methods [217]	Causal methods help identify dependencies between sensitive and non-sensitive variables, making them valuable for detecting proxy biases and conducting subgroup analyses to reveal hidden biases. They enhance transparency in classification-based decision-making by visually representing fairness (or unfairness) within datasets. Directed acyclic graphs (DAGs) are commonly used to illustrate causal relationships, particularly in game theory applications. Additionally, causal methods improve training data quality by inserting, modifying, or removing samples to enforce fairness constraints while maintaining conditional independence within the dataset.	Helps differentiate correlation from causation in AV decision-making, allowing more precise bias mitigation. However, defining correct causal relationships requires domain expertise, and incorrect causal structures can lead to ineffective or misleading bias corrections.
Relabeling [218]	Relabeling (Massaging) adjusts data by reassigning labels to target variables, promoting fairness by elevating disadvantaged groups and reducing the advantage of privileged groups. A ranker is first trained on the original dataset to assign rankings. Then, top-ranked disadvantaged observations initially labeled negatively are reclassified as positive, while bottom-ranked privileged observations are downgraded from positive to negative, ensuring an equal proportion of positive class values across subgroups. However, relabeling has limitations: it is intrusive, as it directly modifies outcome labels, and is restricted to binary-protected attributes, reducing its applicability. Additionally, it defines fairness narrowly by focusing solely on uniform positive class distribution, which may not align with complex real world fairness needs where subgroup differences naturally exist.	Balances training datasets by correcting mislabeled or underrepresented data points. This improves fairness in AV decision-making, but improper relabeling may introduce new biases, and human intervention is often needed, increasing labor costs and the risk of subjectivity.
Perturbation [218]	Perturbation techniques are utilized to detect proxy variables and assess variable influence, and are frequently applied as a preprocessing step prior to regularization, optimization, or reweighting in in-processing approaches. These methods modify data to improve fairness, typically with minimal adverse effects on accuracy. Perturbation can also serve as a means of privacy protection. However, legal constraints may limit the use of data modification methods such as perturbation or relabeling, requiring only minimal changes to ensure compliance. While some classifiers are robust to biases or data modifications, others may degrade in performance when training data is altered for bias mitigation.	Introduces controlled noise to data (e.g., altering lighting conditions in images) to reduce environmental bias in AV perception models. While this enhances generalization, excessive perturbation can degrade model accuracy in normal driving conditions, affecting performance stability.
Variable Blinding [218]	Variable Blinding reduces the influence of sensitive variables in classification models by ensuring the classifier does not differentiate between groups defined by these variables. One approach sets equal threshold values across all groups, effectively removing bias from model outcomes. For example, in loan approvals, group blindness ensures consistent decisions regardless of race or gender. However, another method—omitting sensitive variables from training data—may reduce model accuracy and fail to eliminate discrimination. Both approaches risk overlooking proxy variables, which can still introduce bias. Since discrimination often arises from combinations of variables, careful analysis is essential to ensure effective bias mitigation.	Prevents AV models from relying on biased variables (e.g., ignoring car color if it skews risk assessments). This improves fairness but risks removing essential contextual information, potentially leading to poor decision-making in critical scenarios.
Reweighing [219]	Reweighing approach aims to address bias in training data by assigning weights to individual instances rather than altering the data itself, as seen in other methods like relabeling, perturbation, and transformation. This technique involves finding weights that align the distribution of specific variables between the original and target distributions. By carefully assigning these weights to tuples in the training dataset, the approach ensures fairness in training without modifying any labels, thus eliminating discrimination associated with sensitive variables.	Assigns different weights to underrepresented data samples, ensuring fairer model predictions. It is effective without modifying raw data, but improper weighting can distort decision outcomes, making AV responses less optimal in specific scenarios.

Table 4.8: Static, non-real-time, and reactive post-deployment bias mitigation techniques

Technique	Description	Applicability
Calibrated Odds [220]	Calibrated Odds manages disparities in misclassification costs to reduce discriminatory predictions, even after score calibration across groups. Based on the equalized odds principle, it ensures similar FPR and false-negative rate (FNR) across demographic groups, preventing disproportionate errors. The approach builds an optimal non-discriminating Bayes classifier, considering the joint distribution of true targets, predictions, and protected attributes. It uses randomized information withholding to minimize expected loss while enforcing fairness constraints. However, balancing non-discrimination and calibrated probability estimates is inherently challenging, so the method prioritizes maintaining calibration under a single cost constraint. Its effectiveness depends on the classifier's strong performance, making it less reliable if the underlying model struggles with classification accuracy.	Adjust model predictions to equalize FPR and FNR across demographic or environmental groups. While this helps reduce discriminatory errors, it may slightly reduce overall model accuracy, requiring careful calibration.
Thresholding [220]	Thresholding methods address bias by recognizing that biased decisions often occur near decision boundaries, influenced by human judgment and threshold-based decision rules. These methods identify regions in a classifier's probability distribution where both favored and protected groups receive positive and negative classifications, flagging ambiguous cases as potential bias sources. Researchers use techniques like equalized odds to adjust thresholds for different groups, balancing true and FPR while minimizing classifier loss, promoting fair and accurate performance. However, determining an accepDP trade-off between fairness and accuracy is complex, especially in imbalanced datasets. While thresholding can function as a human-in-the-loop mechanism, improper training may introduce new biases, as fairness is not a simple monotone function, making threshold selection potentially arbitrary. Ultimately, thresholding achieves fairness only when applied judiciously, ensuring equity without compromising accuracy.	Dynamically adjusts AV decision thresholds to correct biases, such as increasing pedestrian detection sensitivity at night. Although effective, it requires continuous monitoring and recalibration to maintain optimal performance across different conditions.
Reject Option Classification [218]	Reject Option Classification enhances fairness by adjusting decisions where the model has low confidence or where predictions fall near the decision boundary. It first identifies ambiguous cases prone to bias, then modifies these decisions to favor disadvantaged groups, improving fairness without significantly compromising accuracy. In some cases, the model may reject the original decision and reclassify instances, even flipping them to the opposite class if it reduces bias. This technique can be optimized for various fairness metrics, such as demographic parity, equalized odds, or disparate impact, making it adaptable across different models. While it effectively mitigates bias and supports multiple fairness goals, it comes with challenges, including a trade-off between overall accuracy and the complexity of fine-tuning and implementation.	Allows AV models to defer uncertain or potentially biased decisions for additional review, preventing unsafe or unfair outcomes. However, this can introduce delays, making it less suiDP for real-time AV decision-making.
Predictive Parity [221]	Predictive Parity is a post-processing technique that adjusts probability thresholds in predictive models to ensure fairness across subgroups defined by protected attributes. It is designed to minimize costs or maximize net benefits while maintaining fairness constraints. Unlike Calibrated Odds, which focuses on FNR and FPR, Predictive Parity allows users to define cost quantifications for misclassification and balances fairness metrics like FPR, TPR, or PPV. This approach is highly versatile, integrating seamlessly into any model workflow without requiring access to predictor features. However, it has limitations—if predictor features correlate with protected attributes, fairness issues may persist, and achieving parity in some metrics may create imbalances in others.	Ensures AV models maintain equal predictive accuracy across different demographic or environmental conditions. While this enhances fairness, it can sometimes reduce overall precision, leading to trade-offs between fairness and model performance.

4.2.3 Post Model deployment - AI Model Monitoring Level Mitigations

Fairness should be continuously monitored to prevent disparate impacts across demographic groups even after the deployment of the AI system. There are two categories in which post-deployment bias mitigation could be classified. The first category listed by DP. 4.8 includes techniques that focus on adjusting decision rules, probability thresholds, or classification confidence levels after deployment but do not inherently adapt to new data in real-time. They rely on fixed fairness constraints defined at deployment and may require manual retraining or parameter tuning to remain effective. Typically, these techniques operate in a batch manner where fairness constraints or decision adjustments are applied after a fixed evaluation period (e.g., periodic model audits or fairness recalibration). Consequently, these techniques are reactive, applied after biases have been detected through audits or fairness checks, and don't provide real-time bias mitigation. The second category of techniques listed by DP. 4.9 is real-time, which proactively adjusts AV decision-making before biases cause safety or fairness issues (e.g., XAI can detect and adjust feature attribution biases before a misclassification occurs, and Online RL refines AV behavior dynamically by modifying its decision-making policies based on real world interactions). These methods are dynamic and continuously evolving, adapting to new environmental conditions, user behaviors, or demographic shifts, and leverage continuous feedback, real-time data streams, and automated adjustments to mitigate bias without human intervention.

It is worth emphasizing the importance of domain adaptation techniques that could play a key role in post-deployment bias mitigation by serving as a powerful tool for bridging the gap between training and deployment environments and ensuring consistent model performance. These techniques could reduce geographic and environmental bias, allowing AVs to operate safely across diverse global locations, minimize costly retraining efforts, enable faster adaptation to new deployment areas with minimal labeled data, and enhance fairness, ensuring equal performance across demographic, environmental, and road variations. The following list provides these techniques that bring reliability

Table 4.9: Dynamic, real-time, and proactive post-deployment bias mitigation during AV operations

Technique	Description	Applicability
Real-Time Fairness Monitoring [222]	Real-Time Fairness Monitoring enables AV systems to continuously detect and correct biases by analyzing how decision outcomes vary across demographic, geographic, or environmental factors, ensuring equiDP behavior for all users. This technique follows a three-step process: (1) Defining fairness metrics, such as equal pedestrian treatment across demographics or consistent behavior in different weather conditions; (2) Monitoring bias in real-time, tracking AV actions like braking, lane changes, and pedestrian yield across various factors (e.g., location, weather, population demographics); and (3) Triggering alerts and dynamically adjusting model parameters when significant bias is detected, such as the AV consistently slowing down less for pedestrians in darker-skinned areas or poorly lit conditions. This approach is particularly effective in cases where pedestrian detection systems fail in low-light environments or with specific clothing colors, allowing real-time sensitivity adjustments to reduce misdetections and enhance safety.	Continuously tracks fairness metrics in AV operations, triggering alerts for bias violations. This enables immediate interventions but demands high computational power, which may impact system efficiency.
Continual Learning for Adaptive Bias Mitigation [223]	Continual Learning enables AVs to adapt in real-time, mitigating biases caused by changing weather, lighting, or traffic conditions. This technique is essential for handling data distribution shifts, which can lead to model drift and reduced fairness. Instead of full retraining, AV models receive incremental updates, incorporating data from new environments to reduce drift bias. Model weights are adjusted dynamically to improve accuracy in underrepresented conditions (e.g., rural vs. urban driving). By identifying high-uncertainty scenarios, continual learning prioritizes improvements where AV models struggle most. For instance, if an AV encounters a newly constructed roundabout with unusual signage, continual learning allows it to generalize this experience to similar future cases, preventing geographical bias where the AV only performs well on standard roundabouts.	Allows AV models to update continuously based on new data, preventing bias accumulation over time. However, if not carefully managed, it risks catastrophic forgetting of previously learned knowledge.
Online Reinforcement Learning (RL) for Real-Time Adaptation [224]	Online Reinforcement Learning (RL) enables AVs to dynamically learn optimal behaviors by interacting with the environment and receiving real-time feedback on decisions. RL algorithms continuously update policies based on environmental changes or observed biases, ensuring fairer decision-making. This is achieved through reward shaping for fairness, where the reward function penalizes biased outcomes, such as failing to stop consistently for all pedestrians. RL also optimizes decision policies by exploring different actions, such as adjusting braking distances within a safety-constrained environment. For example, if pedestrians begin crossing at non-standard points due to construction, RL enables the AV to adapt its pedestrian interaction policy, learning in real-time to yield appropriately, improving safety without bias toward designated crosswalks.	Dynamically refines AV decision policies based on fairness feedback. While this self-correcting mechanism improves fairness over time, real-time deployment is challenging due to potential risks in live training.
Dynamic Threshold Adjustment based on Environment Sensing [225]	Dynamic Threshold Adjustment utilizes real-time sensor data to modify AV decision thresholds based on contextual factors like weather, road type, and pedestrian density, reducing biases related to environmental conditions. Using context-aware sensing, AVs dynamically adjust thresholds for stopping distances, speed, or sensor sensitivity based on learned patterns. For example, if rain or low-light conditions lead to higher false negatives in pedestrian detection, the system increases sensitivity accordingly. Additionally, fairness overrides ensure equiDP treatment in known bias conditions, such as low-visibility scenarios. This technique is particularly effective when AVs struggle with detection in heavy rain or fog, dynamically adjusting thresholds to prioritize safety and prevent bias against vulnerable road users in adverse weather conditions.	Modifies AV decision confidence based on factors like weather or lighting. This ensures consistency in performance but requires real-time environmental sensors and computational power for continuous adjustments.
eXplainable AI (XAI) for Real-Time Bias Transparency and Adjustment [226]	XAI for Real-Time Bias Transparency enables AVs to generate explanations for their decisions, revealing biased decision pathways and providing insights into how and why certain choices were made. This allows for real-time bias mitigation by adjusting decision-making processes accordingly. XAI employs feature attribution monitoring, tracking influential input features (e.g., weather, road type) to detect disproportionate reliance on sensitive attributes, which may indicate bias. If biased features significantly influence decisions, the system alerts AV operators and triggers model adjustments, rebalancing feature weights or decision thresholds. This is particularly useful when pedestrian detection relies too heavily on environmental lighting, potentially causing biases in low-light conditions. In such cases, XAI can trigger real-time adjustments, shifting reliance to alternative features like object motion or shape, ensuring fairer, more balanced decision-making in diverse environments.	Provides interpretability for AV decision-making, allowing real-time bias corrections. While improving accountability and trust, complex XAI methods can slow down decision processes, making real-time applications more difficult.
Federated Learning for Collaborative Bias Mitigation in Real-Time [227]	Federated Learning for Collaborative Bias Mitigation enables AVs to share knowledge across regions without exchanging raw data, allowing biases discovered in one location to inform improvements elsewhere while preserving privacy. Each AV trains locally on real-time, environment-specific data, and periodically shares model updates with the fleet to create a collectively refined model. This ensures cross-region bias mitigation, helping AVs adapt to diverse operational conditions such as differences in pedestrian behavior or road infrastructure while preventing location-specific biases. For example, if AVs in suburban areas frequently encounter pedestrians crossing outside designated zones, federated learning allows urban AVs to adjust their behavior accordingly, ensuring safer and more generalized pedestrian interactions across different regions.	Enables AV fleets to share fairness insights without transferring raw data, ensuring improved bias mitigation while preserving privacy. However, it requires high-bandwidth communication and robust security protocols for effective implementation.

and safety to more RAI-based AV operations.

1) Transfer Learning for efficient model adaptation [228], allows AV models trained in one environment (source domain) to be fine-tuned for better performance in a new, unseen environment (target domain) with minimal labeled data. This functionality is achieved using feature reuse by transferring common driving knowledge (e.g., vehicle detection, lane recognition) from one region to another while adapting to new conditions (e.g., different signage, road textures). To fully unleash the power of transfer learning, fine-tuning with a small amount of labeled data from the deployment environment is performed to adjust key model parameters and model warm-starting, requiring fewer training iterations to adapt to new conditions. This technique is useful when an AV model trained in sunny streets of a warm city can be fine-tuned with limited labeled data from snowy roads to improve perception and decision-making in winter conditions.

2) Unsupervised Domain Adaptation (UDA) leveraged for training with unlabeled data [229] enables model adaptation without requiring labeled data from the deployment environment. In this technique, models learn domain-invariant features that generalize across different locations and are useful when common road features (e.g., road edges, vehicle outlines) are shared across training and deployment conditions. This technique leverages adversarial training for feature alignment where a domain discriminator makes source and target domain features indistinguishable, ensuring smooth adaptation and enabling self-supervised learning on unlabeled data leveraging pseudo-labeling to create synthetic labels in the deployment domain, reducing reliance on expensive data annotation. This technique benefits a self-driving car trained in daylight conditions and can use UDA to adapt to night-time driving without requiring labeled night-time datasets.

3) Few-Shot Domain Adaptation [230] enables AV models to quickly adapt to new deployment environments using only a few labeled examples from the target domain. This technique employs meta-learning for rapid adaptation by leveraging models trained to rapidly generalize when exposed to new environments with minimal labeled data and feature embedding alignment to ensure AV perception models understand the semantic meaning of new features (e.g., different road signs across

different countries). This technique is useful when an AV moving from a country with a right-hand drive to one with a left-hand drive can quickly adapt using just a few labeled images of road signs and lane markings of the target location.

4) Online Domain Adaptation for continuous learning in deployment [231] includes techniques like online RL discussed in DP. 4.9 enabling AVs to adapt in real-time as they collect new data in the deployment environment. This functionality is achieved through incremental model updates, where the AV refines its model while driving, reducing deployment bias as it encounters new road conditions, infrastructure, and weather. It could employ approaches like Bayesian learning to update model parameters dynamically and prioritize high-uncertainty situations for adaptation. Further, it could also leverage active learning with human supervision to identify edge cases (e.g., unusual intersections) and request manual labeling to improve adaptation. This functionality is especially useful for an AV deployed in a newly developed city to continuously update its model to recognize new traffic patterns, road markings, and pedestrian behaviors.

4.2.4 Role and importance of automated tools for continuous fairness auditing in AV

Automated fairness auditing tools provide real-time detection, diagnosis, and mitigation of biases, ensuring AV systems remain fair and equiDP across different deployment environments. The automated fairness auditing tools help implement the real-time bias mitigation techniques and provide the benefits of real-time bias detection, continuous model performance monitoring to ensure AV decision-making remains consistent and unbiased over time, and provide immediate alerts for corrective actions when biases emerge in perception, prediction, or navigation. When seamlessly integrated into AV control systems, they could work within the AV software stack without disrupting AV operations. The following is a list of the core functions of an automated fairness auditing tool that are essential for the future of equiDP, responsible, and transparent AI-driven transportation and

their applicability for ensuring safe, unbiased AV deployment in dynamic environments.

1) Real-time fairness metric monitoring to ensure fairness across all sensor inputs (LiDAR, camera, radar, GPS), prevent biased interpretations, and continuously measure fairness-related performance indicators, such as false positive/false negative rates across different demographic groups, accuracy disparities in object detection (e.g., pedestrians vs. cyclists), and decision consistency across different geographic and environmental conditions. It is important for scenarios where an AV's pedestrian detection system could be less accurate for individuals wearing darker clothing at night. Automated fairness auditing detects this bias in real-time, triggering a model recalibration to improve nighttime pedestrian detection.

2) Automated bias detection and alerts to identify patterns of biased behavior in AV decision-making and trigger alerts when threshold violations occur, such as when AVs brake more aggressively for specific demographics and when lane-change decisions favor one type of vehicle over another. This functionality can also provide explainability reports to help engineers understand the source of bias, and it is applicable for scenarios when an AV frequently misidentifies pedestrians in wheelchairs. The system could detect this issue and prioritize retraining with balanced data.

3) Adaptive bias mitigation through dynamic model recalibration when fairness violations are detected, implementing adaptive decision thresholds to balance accuracy across underrepresented conditions, and employing reinforcement learning to refine real-time decision-making without human intervention. This function is useful for identifying performance gaps and triggering an adaptive recalibration to improve detection in low visibility. For example, it would apply to an AV fleet deployed in a city with frequent fog conditions, which may struggle to detect cyclists.

4) Root cause analysis and bias attribution could help identify whether bias origination across sensor data, model architecture, or deployment conditions, provide granular insights into AV decision processes to pinpoint unfair outcomes, and use causal inference techniques to differentiate between correlation and causation in biased behaviors. This functionality is useful for scenarios when an AV system prioritizes braking for larger vehicles but not for smaller ones like motorcycles. Root cause

analysis could determine that this bias stems from the underrepresentation of motorcycles in the training dataset, prompting synthetic data augmentation.

5) Integration with AV fleet-wide auditing systems could help ensure that all AVs in a fleet are continuously monitored for fairness compliance. This functionality would require using federated learning to share fairness-related insights across multiple vehicles, improving collective decision-making, and aggregate bias reports across diverse locations to identify global trends in fairness violations. It could be applicable to an AV fleet deployed in one urban setting experiencing fairness violations due to high pedestrian density while another shows biases related to extreme lighting conditions. The system could aggregate insights from both locations to develop a more universally fair model.

4.2.5 Handling Rare and Critical Scenarios via Synthetic Data

The performance of AVs heavily relies on diverse and representative training data. However, real world data collection can sometimes be costly or risky, making data availability a challenge. In other cases, data may lack sufficient coverage of rare but high-risk scenarios such as pedestrian accidents in adverse weather, unexpected road obstacles, or extreme lighting conditions. Synthetic data generation provides a powerful solution by artificially creating diverse driving scenarios underrepresented in real world datasets, such as by simulating edge case scenarios that are unlikely to be captured in standard datasets or by helping minimize data imbalance bias by ensuring fair representation of different road users, environments, and rare events. It also helps accelerate AV model training and testing efforts by enabling training and validation under extreme conditions before real world deployment. The following are details on synthetic data generation techniques and their applicability for the AV development lifecycle: **1) Generative Adversarial Networks (GANs)**[232]: GANs use a generator-discriminator framework to create highly realistic synthetic images, videos, and sensor data. GANs simulate adverse weather conditions by transforming existing clear-weather driving scenes into rainy, snowy, or foggy environments, improving AV robustness

across conditions. GANs can also modify real world driving footage to insert synthetic pedestrians at risky locations, such as behind a driver's field of vision or in nighttime conditions, to handle near-miss pedestrian scenarios. Furthermore, GANs can help reduce bias and improve model fairness by augmenting data with underrepresented classes like wheelchair users crossing roads.

2) Variational Autoencoders (VAEs)[233]: VAEs generate diverse but realistic variations of input data, making them useful for edge-case simulation and scenario augmentation. VAEs enable edge case exploration further by allowing AVs to be tested on never-before-seen conditions by interpolating between normal and extreme scenarios. For example, VAEs can generate gradual transitions between clear and foggy driving conditions, helping AV models adapt to changing environments in real-time. Unlike GANs, VAEs can synthesize sensor data (LiDAR and radar), which is essential for multimodal AV perception.

3) Combining GANs/VAEs with Physics-based simulation. This approach leverages a physics-based AV simulator with GANs/VAEs to enhance performance in real world conditions. A physics-based driving simulators ensure synthetic scenarios adhere to real world physics. This approach can significantly improve AV performance by creating a training dataset containing virtual replicas of cities, weather conditions, and road users to simulate AV interactions. It could then leverage GANs/VAEs to generate AI-generated textures, objects, and sensor noise layers and add those to make synthetic data visually and statistically realistic. Further, this approach could be leveraged for automated annotation, where, unlike real world datasets, synthetic data provides perfectly labeled ground truth and accelerates AV model training.

While synthetic data can improve AV performance significantly, it poses specific challenges, such as ensuring realism for training data, avoiding overfitting to synthetic patterns, and scaling up the computational and training costs. Leveraging physics-based simulators with GANs/VAEs can help AVs detect bias in real-time to identify systematic failure cases further, generate targeted synthetic scenarios to correct these biases, and help retrain models dynamically using federated learning with synthetic-augmented data. A real world example could include a scenario where

an AV struggling with snow-covered pedestrian crossings could trigger on-the-fly synthetic data generation, updating its model with new GAN-generated winter pedestrian scenes to handle this scenario.

4.3 Simulation and Results

Leveraging the discussed bias types and applicable mitigations, the RAI Framework presented in Fig. 4.2 is leveraged in this section to evaluate bias mitigations across the AI lifecycle for AV systems. In order to complete the evaluation, the following inputs and considerations are made to perform the simulations to obtain the results and conduct the analysis:

Input Dataset: A diverse driving dataset BDD100K [234] from the University of California, Berkeley, was considered for this analysis, where selected attributes like weather, time of day, and scene are considered in the analysis.

Synthetic Data Generation: Two sets of synthetic data generation are employed during the evaluation. (1) The first set includes appending each record of the BDD100K dataset with demographic data points, i.e., age, gender, and race, along with a binary target value, which are collectively leveraged to analyze the impact of bias. This set included 69,863 in the original dataset and was further appended with columns (age, gender, and race) while keeping the record count the same. The data distribution opted for column 'age' was 20%-60%-20% for child-adult-senior, for column 'gender' was 50% for each male and female, and for column 'race' was 50%-20%-20%-10% for white-black-asian-others type. Please note that all these data distributions are illustrative and could be changed to adapt to different demographic scenarios. (2) The second set of synthetic data is generated to analyze the impact of variations in the demographics, which was achieved by varying the data distribution of the sensitive variable 'gender' to 70%-30% for male-female as opposed to the previous distribution of 50% for each. This analysis assumes that one of the two data distributions contains bias with respect to the other since both have different gender ratios while keeping all other

parameters the same.

Sensitive & Target Variables: To analyze the bias, attributes like 'race', 'gender', and 'age' can be considered as the sensitive variable with the goal of understanding whether the model outcome represented by the target variable is impacted by certain demographic groups comprising of different races. It is important to understand whether any social or regional biases have crept into the training dataset and also crucial to ensure regulatory compliance by avoiding non-discriminatory behavior by the AI/ML model. It is important to note that without loss of generality, this approach can also analyze bias impact due to other demographic attributes not included in this dataset. For the simulations in this section, 'gender' is selected as the 'sensitive' variable while a dependent 'target' variable to analyze the bias impact was generated with equal outcome distribution of 50% each for positive and negative outcomes.

Fairness Metrics for Bias Evaluation: This study considered the following most commonly used fairness metrics [235] to assess bias mitigation:

(i) Disparate Impact (DI) assesses the ratio of positive outcome rates between groups such as privileged versus unprivileged groups. A DI value close to 1 indicates fairness, meaning the positive outcome rates are proportionate between groups.

(ii) Demographic Parity Difference (DPD) measures the difference in positive classification rates between groups or the difference in the probability of receiving a positive outcome between two groups where a lower DP indicates greater fairness. This could be leveraged to understand the bias while comparing a potentially underprivileged group with a privileged one.

(iii) Equalized Odds Difference (EOD) measures the difference in True Positive Rates (TPR) and False Positive Rates (FPR) between groups and helps evaluate if TPR and FPR are balanced across different groups. Ideally, both metrics should be similar for all groups. In other words, the comparison of EO between the groups should be closer to zero, indicating lesser disparity.

(iv) Fairness in Accuracy (FIA) measures whether an AI model provides comparable accuracy rates across different demographic groups (e.g., race, gender, age). The value of 1 signifies that the

model has equal accuracy for both privileged and unprivileged groups, i.e., no accuracy disparity. On the other hand, the values closer to zero depict maximum disparity, i.e., accuracy for one group is perfect, while for the other, it's at random (or 0%).

(v) Intersectional Fairness Score (IFS) measures fairness by evaluating the average accuracy across different groups formed by combinations of sensitive attributes like race, gender, and age rather than considering them independently. IFS values close to 1 indicate fair accuracy across most groups, while values lower than 0.5 typically highlight some intersectional groups are experiencing lower accuracy rates. This is an important metric as it reveals how different biases may compound when multiple demographics intersect. For example, a model may score high on FIA (indicating fairness across single attributes) but still have low IFS, suggesting that it fails at the intersection of those attributes.

4.3.1 Comparison of AI model pre-design level bias mitigation techniques

In this experiment, this study presents the impact of the commonly used data sampling techniques - stratified sampling and adaptive sampling, along with Adversarial Debiasing and Regularization techniques - and evaluated their impact on bias mitigation and fairness in the ML model for AVs. The analysis evaluates and compares fairness metrics before and after applying each sampling technique to assess its impact on bias mitigation, and the results are highlighted in Fig. 4.3.

Analyzing Disparate Impact, it can be noticed that all three mitigation techniques, except Stratified sampling, resulted in poor performance and made the bias impact larger than the baseline model. Typically, DI values over 1.25 and under 0.8 signify unfair treatment between the compared groups, highlighting the poor performance of the three approaches in this case. However, applying Stratified sampling not only resulted in significantly improving the DI, it demonstrated a slight improvement in model accuracy as well. For DPD, EOD, FA, and IFS, Stratified sampling further outperformed the other techniques. It is important to note that applying adversarial debiasing resulted in making the AI model more discriminatory than the baseline, signified by the negative

Figure 4.3: Fairness Metrics Report for the Analyzed AV Dataset through different Model Pre-Design Level Bias Mitigation Techniques

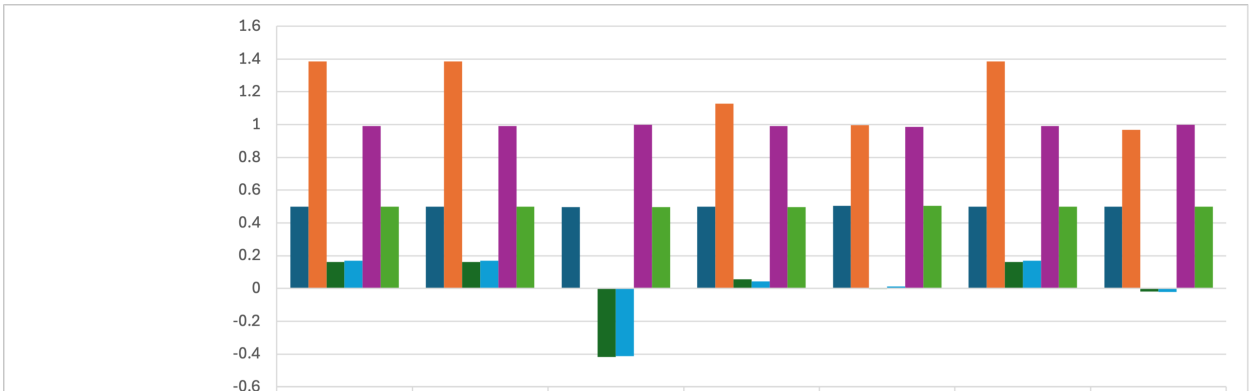
values. Therefore, in this case, the underprivileged group is less likely to obtain positive outcomes compared to the privileged group.

Another important highlight is to compare the performance of the Stratified Sampling for the BDD100K dataset against a synthetically generated dataset where the data distribution of the sensitive variable was slightly altered to analyze its impact on the overall performance. Specifically, the gender ratio between the two datasets was altered from "female = 34970 versus male = 34893" to "female = 34789 versus male = 35074". It could be noticed that this resulted in a significant performance improvement across all the techniques and, more specifically, for Stratified sampling. This could prove to be a very useful application when improving data collection techniques at the

source to ensure the data distribution of the sampling does not result in the introduction of bias. In other cases, it could work as feedback to improve additional data collection on an existing process.

By comparing results across the four debiasing techniques, it can be observed that the stratified sampling technique outperformed all other techniques for this specific dataset while increasing the overall accuracy of the model when compared against the baseline numbers. Therefore, for this dataset, the framework automatically selects Stratified sampling while applying debiasing techniques during the pre-AI model design stage to achieve higher fairness in the outcomes. It is important to highlight that the framework's selection of the specific sampling technique is based on the chosen dataset and could be generalized for any dataset, which is an important contribution of this research.

4.3.2 Comparison of AI model design, development, & deployment level bias mitigation techniques



	Baseline Model	Reweighing	Relabeling	Perturbation	Variable Blinding	Causal Learning	Adversarial Learning
Accuracy	0.5	0.5	0.497	0.499	0.503	0.5	0.498
Disparate Impact	1.385	1.385	0	1.128	0.997	1.385	0.968
Demographic Parity Difference	0.161	0.161	-0.417	0.057	-0.002	0.161	-0.019
Equalized Odds Difference	0.17	0.17	-0.413	0.043	0.0126	0.17	-0.022
Fairness in Accuracy	0.99	0.99	0.999	0.99	0.986	0.99	0.999
Intersectional Fairness Score	0.5	0.5	0.497	0.497	0.503	0.5	0.498

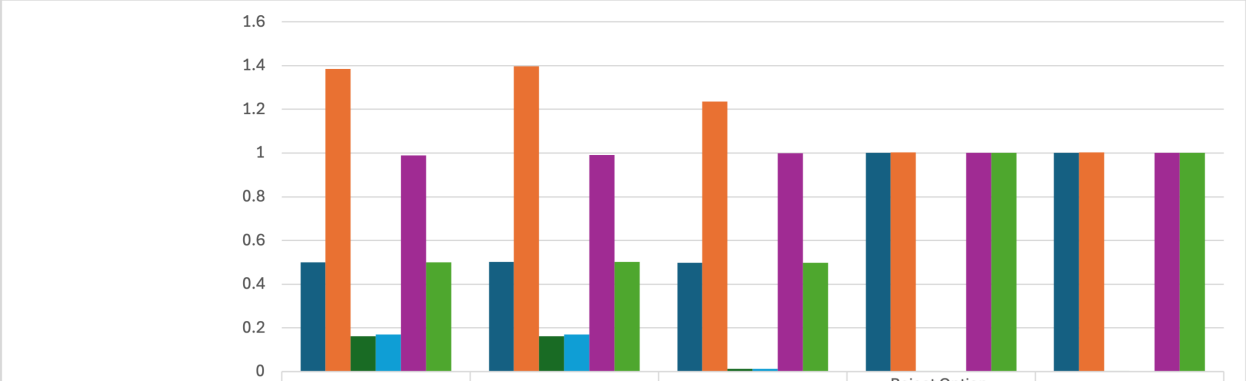
Figure 4.4: Fairness Metrics Report for the Analyzed AV Dataset through different Model Design and Development Level Bias Mitigation Techniques

This experiment involved simulating the model behavior while leveraging the different model

design and development stage bias mitigation techniques. It can be noticed from the results presented in Fig. 4.4 that reweighing had little to no impact on the model performance overall when compared across accuracy and different fairness metrics. On the other hand, Relabeling, Variable Blinding, and Adversarial Learning resulted in negative values, highlighting the increased bias in making one gender more underprivileged than the other one. Through collective comparison across all the fairness metrics, it could be noticed that Perturbation outperformed the other techniques and was selected by the framework for this specific dataset. The model design and development level bias mitigation plays a vital role where altering the data is not possible, such as in cases where data collection and processing are completed and can't be managed, and bias reduction can only be handled at later stages, such as model design and development. In other cases, this approach can be leveraged where AI models are pre-trained on biased datasets and require fine-tuning on target datasets with different biases or distributions.

4.3.3 Comparison of post-AI model deployment level mitigation techniques

This experiment involved simulating the model behavior post-deployment, leveraging the discussed bias mitigation techniques. It can be noticed from the results presented in Fig. 4.5 that both the Reject Option Classification and Predictive Parity Techniques demonstrated similar behavior and outperformed the other two techniques, ensuring selection by the framework for this dataset. Also, it is important to highlight that while applying the thresholding technique improved the DI over the base model, it had a negative impact on the model's accuracy. This is an important highlight as in some cases where there could be technical challenges in applying different mitigations, the limitations of an individual technique should be weighed against its benefits during its selection for bias mitigation for a specific dataset and model. Please note that this set of post-AI model deployment techniques plays a vital role when third-party AI models are leveraged, and dataset-level



	Baseline Model	Calibrated Odds	Thresholding	Reject Option Classification	Predictive Parity
Accuracy	0.5	0.501	0.498	1	1
Disparate Impact	1.385	1.396	1.236	1.003	1.003
Demographic Parity Difference	0.161	0.162	0.013	0.001	0.001
Equalized Odds Difference	0.17	0.17	0.013	0	0
Fairness in Accuracy	0.99	0.992	0.999	1	1
Intersectional Fairness Score	0.5	0.501	0.498	1	1

Figure 4.5: Fairness Metrics Report for the Analyzed AV Dataset through different Post Model Deployment Bias Mitigation Techniques

mitigations could still play a role in improving model performance or preventing it from becoming worse due to the discussed post-deployment risks.

4.3.4 Importance of IFS for AI-based AV operations

Ensuring intersectional fairness in AI-based AVs is essential to prevent models from disproportionately misjudging specific demographic combinations across age, gender, and race, which may result in significant real-world consequences for passenger and pedestrian safety. Although traditional bias mitigation approaches may address fairness for single variables, such as race or gender, intersectional biases necessitate more sophisticated strategies to address the complex and overlapping disparities that persist after deployment. Simulation results for the current dataset indicate that the IFS across all mitigation techniques was approximately 0.5, reflecting poor performance across demographic combinations. This outcome may be attributed to one or more of the following factors: **1)** Imbalanced representation across intersectional subgroups can occur even when individual attributes, such as gender or race, appear balanced. Certain combinations, for example, senior-other-

female, may be significantly underrepresented. This underrepresentation can result in poor model performance for these smaller subgroups, thereby negatively affecting IFS.

2) As discussed in Section 4.1.4, although mitigation strategies may address biases associated with individual attributes such as race or gender, intersectional biases arising from the interaction of multiple attributes can remain. For instance, a model may achieve fairness for senior and female groups separately, yet still misclassify senior women at a higher rate due to the compounding effects of overlapping biases.

3) Limited model complexity or inadequate training data can lead to overfitting on dominant groups within the training dataset, while failing to generalize to less common combinations. This results in reduced accuracy and fairness for underrepresented demographics.

The adoption and implementation of holistic data strategies, intersection-aware algorithms, and fine-grained evaluations are critical for ensuring that AV systems make fair and reliable decisions across all demographic groups, thereby advancing the IFS toward a value of 1. Potential approaches include:

1) Increase intersectional representation by employing adaptive and stratified sampling, as well as synthetic data generation, to address underrepresented demographic combinations.

2) Implement intersection-aware mitigation strategies, such as intersectional reweighing and conditional demographic parity, to address biases across multiple attributes simultaneously.

3) Enhance model complexity through transfer learning and fine-tuning to improve generalization to rare demographic groups.

4) Conduct fine-grained fairness evaluations using confusion matrices and heatmaps for all intersectional subgroups to enable targeted interventions by identifying demographic combinations that contribute to disparities.

Collectively, these techniques may advance the IFS toward 1.0, thereby ensuring that AI systems in AVs make fair and reliable decisions across all demographic intersections.

4.4 Summary

In conclusion, this chapter conceptualized bias and fairness as risks that span the entire lifecycle of AI-based AVs. Disparities may be introduced prior to model design, for example, through skewed data and labelling, and can be amplified during model development and deployment due to objective selection and performance tradeoffs. These disparities may also persist after deployment in the absence of continuous monitoring. The chapter maps these risks to practical mitigation strategies at the data, model, and post-deployment stages, emphasizing the importance of automated, ongoing fairness auditing and the use of synthetic data to address rare but safety-critical scenarios. Simulation-based comparisons across multiple fairness and accuracy measures, including intersectional evaluation, demonstrate that aggregate performance metrics can obscure significant subgroup and intersectional harms. Improving equity for underrepresented combinations of sensitive attributes requires intentional representation, intersection-aware mitigation, and detailed evaluation. The chapter concludes that robust AV deployment relies on comprehensive data strategies, intersection-aware algorithms, and continuous, governance-driven monitoring to ensure fair and reliable decisions for all demographic groups.

5 | A Responsible Generative Artificial Intelligence based Multi-Agent Framework for Preserving Data Utility and Privacy

5.1 Proposed Agentic AI Framework

This study introduces a hierarchical, multi-agent framework for adaptive privacy-preserving data exchange that integrates human-centric utility reasoning with automated differential privacy optimization. The framework comprises two principal agents: the Adaptive Utility Reasoning Agent (AURA) and the Modular Agentic Engine for Strategic Tuning & Reporting Orchestration agent (MAESTRO), each supported by four specialized sub-agents, as illustrated in Figure 5.1. Together, these agents facilitate five primary research contributions: adaptive qualitative–quantitative translation, context-aware optimization, feature-driven utility discovery, dynamic privacy-utility balancing, and adaptive intelligence through continuous learning.

AURA is responsible for utility cognition by converting subjective user intent into interpretable and adaptive utility scores, while MAESTRO manages privacy actuation by translating utility objectives into calibrated differential privacy mechanisms. The outputs generated by AURA serve

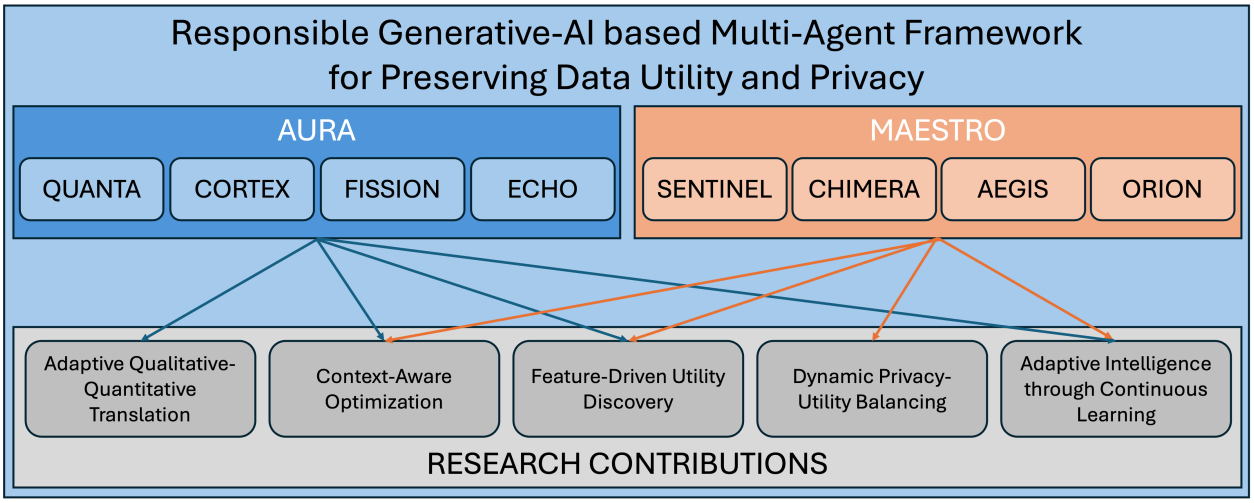


Figure 5.1: Proposed Framework highlighting its constituent Master and Sub-Agents

as inputs for MAESTRO, reflecting the hierarchical structure of the framework. Both AURA and MAESTRO achieve their objectives iteratively through their respective sub-agents. The framework operates according to a sense–reason–act–learn paradigm: it senses qualitative intent and dataset characteristics, reasons over context and feature relevance, acts through privacy parameter selection and data augmentation, and learns continuously from outcomes and feedback. This division of responsibilities enables modular extensibility, interpretability, and robustness across diverse datasets and evolving user requirements.

5.1.1 AURA and the Sub-Agent Description

AURA enables transforming users' qualitative assessments of data usefulness into quantitative, adaptive metrics. Supported by several specialized sub-agents, AURA operates as the primary orchestrator, dynamically evaluating data utility through coordinated agent interactions. It converts subjective descriptors such as "high completeness," "low uniqueness," and "medium diversity" into numerical scores, adjusts weighting schemes according to dataset context, identifies the most influential factors, and refines predictions over time through continuous feedback. By integrating user input and maintaining learned mappings, thresholds, and contextual models, AURA functions

as a meta-cognitive controller that balances interpretability and adaptability in estimating data utility across varied datasets. Table 5.1 details the functionality of each sub-agent supporting AURA.

Table 5.1: Sub-Agents supporting AURA's functionality.

Sub-Agent Name	**Core Function**
Qualitative-qUANtitative Translation Agent (QUANTA)	Learns how users' qualitative judgments, such as "high completeness", map onto a numeric range (0–1). Updated continuously through supervised or feedback-driven learning, it serves as the bridge between human perception and measurable data metrics.
Context-Oriented Reasoning & Threshold-Extraction eXpert agent (CORTEX)	Derives optimal weightings for each attribute by considering the dataset's context (size, structure, missingness) using its learned patterns. Acting as AURA's analytical core, it dynamically adjusts weight distributions.
Feature Importance & Selection Intelligence ON-agent (FISSION)	Discerns which data attributes most strongly influence perceived utility, employing mutual-information or variance-based heuristics. Its role is to separate signal from noise and pinpoint the true drivers of utility.
Experience-Consolidating Heuristic Optimizer agent (ECHO)	Handles ongoing learning by recording outcomes, user validations, and internal states, which are fed back into all models. ECHO functions as AURA's long-term memory and feedback resonance system.

5.1.2 MAESTRO and the Sub-Agent Description

Within the agentic framework, MAESTRO extends AURA by incorporating a target-utility objective and an augmentation loop. It establishes a hierarchical system that dynamically learns and adjusts differential privacy parameters to achieve an optimal balance between data privacy and

semantic utility. MAESTRO enforces user-defined data utility targets while maintaining differential privacy across heterogeneous textual datasets. The system comprises four sub-agents, as outlined in Table 5.2. In addition to coordinating these sub-agents, as illustrated in Figure 5.1, MAESTRO executes augmentation alpha-loops, updates policy learning files, and manages convergence toward the user-specified target utility.

Table 5.2: Sub-Agents supporting MAESTRO's functionality.

Sub-Agent Name	**Core Function**
Semantic & sENsitivity Text INspector for Embedding aNaLytics agent (SENTINEL)	It acts as the profiler agent to profile sensitive text columns, estimates sensitivity (S) and semantic density (ρ), and generates embeddings.
Corpus Hallucination & Iterative Model for Enhanced Resemblance & Anonymity agent (CHIMERA)	As a synthesis agent, it synthesizes augmented textual samples using generative modeling or n-gram fallback to enhance representational diversity.
Adaptive Epsilon & Gaussian/Laplace Intelligent Selector agent (AEGIS)	It serves as the privacy agent and performs the selection of optimal noise mechanisms (Laplace or Gaussian) and tunes ε, δ, and σ dynamically using S and ρ.
Objective Retention & Information Optimization Nexus agent (ORION)	It works as the utility agent that quantifies semantic and structural utility degradation using cosine similarity and silhouette metrics.

5.2 Framework Implementation Methodology

The current implementation of the proposed agentic, adaptive framework operates on any tabular dataset with a header row. However, without loss of generality, the framework can be easily extended for additional dataset formats beyond tabular data such as XML and JSON. The current

implementation considers the following four core data quality dimensions:

1. Completeness, highlighting the proportion of non-missing cells in the dataset.
2. Uniqueness, denoting the fraction of unique rows relative to total rows.
3. Diversity, signifying the average normalized Shannon entropy across columns, measuring the variety of values.
4. Format Consistency, symbolizing the proportion of values in each column matching the inferred dominant data type (integer, float, datetime, or string).

These data quality dimensions are normalized to the range $[0, 1]$, enabling aggregation and comparison across heterogeneous datasets.

5.2.1 AURA Implementation

AURA coordinates the pipeline, ensuring alignment between qualitative perceptions and quantitative assessments, and consolidates learning signals. These unified signals are disseminated to all sub-agents. The four autonomous agents, integrated by AURA, collectively map subjective labels to numeric values, adjust weightings based on dataset context, identify principal utility drivers, and improve performance through iterative feedback. AURA defines data utility as a function of intrinsic dataset characteristics, contextual signals, and learned mappings that align subjective human perception with quantifiable evidence.

Formally, for a dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n denotes number of records (rows) in the dataset and d denotes number of attributes (columns) in the dataset, AURA specifies a vector of normalized characteristic scores as:

$$\mathbf{s} = (s_{\text{comp}}, s_{\text{uniq}}, s_{\text{div}}, s_{\text{fmt}}) \tag{5.1}$$

where each $s_{\text{c}} \in [0, 1]$ quantifies a fundamental aspect of data quality:

- **Completeness** denotes the degree to which expected attributes/values are present in a dataset [236]:

$$s_{\text{comp}} = 1 - \frac{\text{missing cells}}{n \times d} \tag{5.2}$$

- **Uniqueness** denotes the absence of duplicated records [237]:

$$s_{\text{uniq}} = \frac{\text{unique rows}}{n} \tag{5.3}$$

- **Diversity**: Let $X_{\cdot j}$ denote column j and $p_j(v)$ the empirical probability of value v in that column. The (normalized) Shannon entropy [238] per column will be

$$H^*(X_{\cdot j}) = \begin{cases} -\dfrac{\sum_v p_j(v)\ \log p_j(v)}{\log\left(|\text{supp}(X_{\cdot j})|\right)} & \text{if } |\text{supp}(X_{\cdot j})| \geqslant 2, \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

and the dataset-level diversity will be

$$s_{\text{div}} = \frac{1}{d} \sum_{j=1}^{d} H^*(X_{\cdot j}) \tag{5.5}$$

where, $\mathbf{X}$: dataset; $X_{\cdot j}$: column j; v: a distinct observed value in column j; $\text{supp}(X_{\cdot j})$: set of distinct observed values in column j; $|\text{supp}(\cdot)|$: number of distinct observed values; $p_j(v)$: empirical probability of value v in column j; $\log(\cdot)$: logarithm (any fixed base); $H(\cdot)$: Shannon entropy; $H^*(\cdot) \in [0, 1]$: normalized entropy (entropy divided by maximum $\log(|\text{supp}|)$); d: number of columns; $s_{\text{div}}(\mathbf{X}) \in [0, 1]$: average normalized diversity across columns.

- **Format Consistency** measures the conformity of format of the same data in different places in the dataset [239]. Let $\pi_j \in \{\text{int}, \text{float}, \text{datetime}, \text{string}\}$ be the dominant inferred type in column j. Define an indicator $I_{\pi_j}(x) = 1$ if x matches π_j, else 0. Then

$$s_{\text{fmt}}(\mathbf{X}) = \frac{1}{d} \sum_{j=1}^{d} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_{ij} \in \mathcal{F}_j\} \right), \tag{5.6}$$

where, $\mathbf{X}$: dataset; n: number of rows; d: number of columns; X_{ij}: cell at row i, column j; $\mathcal{F}_j$: accepted format specification for column j (e.g., type constraint, regex pattern, standardized date format); $\mathbb{I}\{\cdot\}$: indicator (1 if condition true, else 0); $s_{\text{fmt}}(\mathbf{X}) \in [0, 1]$: proportion of values conforming to the column format, averaged across columns.

The overall data utility is computed by aggregating these characteristics using context-aware weights $w(x)$, which are determined by the dataset profile x:

$$U(\mathbf{X}; x) = \sum_{c \in C} w_c(x)\, s_c, \qquad \sum_{c} w_c(x) = 1 \tag{5.7}$$

This formulation enables AURA to model data quality and utility relationships that differ across domains or scales. For instance, temporal datasets may emphasize diversity, whereas administrative records may prioritize completeness.

AURA, functioning as the master intelligence layer that coordinates all sub-agents, receives user qualitative expectations and raw dataset characteristics, directs QUANTA to convert linguistic ratings into numerical thresholds, instructs CORTEX to infer context-sensitive weights, retrieves FISSION's learned importance scores, and updates its evolving memory through ECHO. AURA subsequently integrates these outputs to compute an overall data utility score for any dataset. The implementation details of the four autonomous agents are as follows:

5.2.1.1 QUANTA Implementation

QUANTA offers a formalized method for translating subjective human language into machine-interpretable numeric constraints, utilizing a self-correcting update mechanism. This methodology introduces a human-aligned interpretative layer within the data utility pipeline that adapts over time by

refining numeric values based on user validation and observed dataset performance. The QUANTA agent converts human descriptions, such as "high completeness" or "moderate uniqueness," into precise numerical thresholds. Qualitative user inputs (very low, low, moderate, high, very high) are mapped to quantitative values by defining an ordered linguistic scale $\mathcal{L} = \{\text{VL}, \text{L}, \text{M}, \text{H}, \text{VH}\}$, with the mapping $m : \mathcal{L} \rightarrow [0, 1]$ assigning each label a numeric threshold $\tau_c = m(q_c)$ for each characteristic c.

Label Mapping by QUANTA enforces an ordered scale, ensuring that “very low” always maps to a smaller numeric value than “low”, “moderate”, and so on, to protect logical consistency in the mapping.

$$m(\text{VL}) < m(\text{L}) < m(\text{M}) < m(\text{H}) < m(\text{VH}) \tag{5.8}$$

Similarly, during threshold assignment, each characteristic c receives a numeric threshold τ_c derived from the qualitative label q_c using QUANTA’s learned mapping function $m(\cdot)$.

$$\tau_c = m(q_c) \tag{5.9}$$

Given a user-specified overall target utility U^* and observed characteristic scores, $\mathbf{s}$, QUANTA predicts an internal proxy utility:

$$\widehat{U} \;=\; \frac{1}{|C_q|} \sum_{c \in C_q} \Big(1 - \big|s_c - \tau_c\big|\Big) \tag{5.10}$$

where, C_q represents the subset of characteristics for which the user provided a qualitative judgment, as opposed to the full set of data-quality characteristics (completeness, uniqueness, diversity, and format).

Algorithm 1: QUANTA: Qualitative to Quantitative Translation

Input: Qualitative labels $\{q_c\}$, characteristic scores $\mathbf{s}$, optional target utility $U^\star$, current mapping $m : \mathcal{L} \to [0, 1]$.

Output: Numeric thresholds $\{\tau_c\}$, updated mapping m.

Step 1: Initialize and Enforce Ordering
Ensure $m(\text{VL}) < m(\text{L}) < m(\text{M}) < m(\text{H}) < m(\text{VH})$
If a label q_c is unseen, initialize $m(q_c)$ to a prior (e.g. 0.5)

Step 2: Compute Thresholds
foreach *characteristic c with label* q_c **do**
  $\tau_c \leftarrow m(q_c)$

Step 3: If Target Utility is provided, adjust Mapping
if $U^\star$ *is provided* **then**
  Compute proxy utility

$$\widehat{U} \;=\; \frac{1}{|C_q|} \sum_{c \in C_q} \Big(1 - \big|s_c - \tau_c\big|\Big)$$

  Compute error $e = U^\star - \widehat{U}$
  foreach $c \in C_q$ **do**
    $m(q_c) \leftarrow \Pi_{[0,1]}\Big(m(q_c) + \eta \,\text{sgn}(e)(s_c - \tau_c)\Big)$
  Apply isotonic projection to restore ordering $m(\text{VL}) < \cdots < m(\text{VH})$
else
  Do nothing (no supervised correction)

return $\{\tau_c\}, m$

If the system's predicted utility $\widehat{U}$ differs from the user's target utility $U^\star$, the mapping adjusts slightly in a direction that reduces this error; $\Pi_{[0,1]}$ ensures the new value remains valid.

$$m(q_c) \;\leftarrow\; \Pi_{[0,1]}\Big(m(q_c) + \eta \,\text{sgn}(U^* - \widehat{U})\,(s_c - \tau_c)\Big) \tag{5.11}$$

This procedure ensures that the qualitative ordering is preserved through isotonic projection. A detailed illustration for the process of QUANTA is provided in Algorithm 1.

5.2.1.2 CORTEX Implementation

CORTEX employs a context-sensitive weighting mechanism that adapts to the structure of each dataset. In contrast to fixed-weight systems, it constructs a functional mapping from contextual features to utility weights, which enables AURA to adjust its behavior across diverse domains such as EV charging logs and demographic data. CORTEX evaluates the relative importance of each data-quality characteristic for a given dataset; for example, when a dataset contains many missing values, it assigns greater weight to completeness, whereas in datasets with high variety, diversity is prioritized. To facilitate this process, the CORTEX agent defines a context-sensitive function $\mathrm{w}(x)$ that assigns relevance weights to each characteristic based on the dataset profile x, which may include factors such as the number of rows, degree of missingness, and average cardinality.

$$w_c(x) \;=\; \frac{\max\{0, \boldsymbol{\theta}_c^\top x + b_c\}}{\sum_{c'} \max\{0, \boldsymbol{\theta}_c^\top x + b_c(x)\}} \tag{5.12}$$

where, $\boldsymbol{\theta}_c$ and b_c parameterize a characteristic-specific linear model that maps dataset context features to the unnormalized importance of characteristic c, enabling context-aware and interpretable weight inference. Here, $\boldsymbol{\theta}_c$ denotes a learned parameter vector that tells CORTEX how strongly each context feature should influence the importance of characteristic c. Whereas, b_c is a bias (intercept) term representing the baseline importance of characteristic c when context features are neutral or zero.

CORTEX employs a regression model to link dataset context features $x^{(t)}$ (e.g., the percentage of missing values and the dataset cardinality) to expert-like weight assignments $w_c^{(t)}$. A regularization term is included to prevent overfitting while CORTEX is trained using ridge regression on historical data:

$$\min_{\boldsymbol{\theta}_c, b_c} \sum_t \left(w_c^{(t)} - (\boldsymbol{\theta}_c^\top x^{(t)} + b_c)\right)^2 + \lambda \|\boldsymbol{\theta}_c\|_2^2 \tag{5.13}$$

Algorithm 2: CORTEX: Context-Aware Weighting

Input: Context vector x, history $\mathcal{H}$, current regression parameters $\{\theta_c, b_c\}_c$.

Output: Weights $\mathbf{w}(x)$ for characteristics C.

Step 1: Check History Sufficiency
Extract training pairs $\{(x^{(t)}, \mathbf{w}^{(t)})\}_{t=1}^{T}$ from $\mathcal{H}$

if *T is large enough* **then**
  Step 2: Supervised Weight Learning
  foreach $c \in C$ **do**
    Solve the ridge regression problem:

$$\min_{\theta_c, b_c} \sum_{t=1}^{T} \left(w_c^{(t)} - (\theta_c^\top x^{(t)} + b_c) \right)^2 + \lambda \|\theta_c\|_2^2$$

else
  Use prior or default parameters for $\{\theta_c, b_c\}_c$

Step 3: Predict Raw Weights
foreach $c \in C$ **do**
  $\widetilde{w}_c(x) \leftarrow \theta_c^\top x + b_c$

Step 4: Normalize onto Simplex
Compute $w_c(x)$ for all c as:

$$w_c(x) = \frac{\max\{0, \widetilde{w}_c(x)\}}{\sum_{c'} \max\{0, \widetilde{w}_{c'}(x)\}}$$

Step 5: Heuristic Fallback (Optional)
If the denominator is zero or the history is too small:
1. Define heuristic scores $h_c(x) \geqslant 0$ (e.g., emphasize completeness if missingness is high)
2. Set

$$w_c(x) = \frac{h_c(x)}{\sum_{c'} h_{c'}(x)}$$

return $\mathbf{w}(x)$

Normalized weight prediction converts the raw outputs $w_c(x)$ into a probability-like distribution, ensuring non-negativity and that the weights sum to one.

$$w_c(x) = \frac{\max\{0, \widetilde{w}_c(x)\}}{\sum_{c'} \max\{0, \widetilde{w}_{c'}(x)\}}, \qquad \widetilde{w}_c(x) = \theta_c^\top x + b_c \tag{5.14}$$

When historical data is insufficient, CORTEX defaults to heuristic priors, such as increasing the completeness weight when missingness is high. In this case, a heuristic $h_c(x) \geqslant 0$ is applied (for example, emphasizing completeness under high missingness):

$$h_{\text{comp}}(x) = 1 + \alpha\,(\%\text{missing}), \quad h_{\text{div}}(x) = 1 + \beta\,(\text{avg. cardinality}), \quad w_c(x) = \frac{h_c(x)}{\sum_{c'} h_{c'}(x)} \tag{5.15}$$

A detailed illustration for the process of CORTEX is provided in Algorithm 2.

5.2.1.3 FISSION Implementation

FISSION supports data-driven identification of influential quality dimensions, enabling AURA to iteratively refine its reasoning and provide explanations based on empirical utility behavior. The FISSION agent identifies dataset characteristics, such as completeness and uniqueness, that most significantly explain changes in utility over time. This methodology improves system interpretability and promotes experiential learning. Feature importances are derived from historical run data $(s^{(t)}, y^{(t)})_{t=1}^{T}$. When feasible, mutual information regression is conducted between characteristic scores and validated overall utility scores; otherwise, a variance-based heuristic is employed. The resulting feature importance scores enhance both interpretability and the learning process of the weighting agent.

When mutual information is available, the importance φ_c increases if changes in the characteristic score s_c strongly predict changes in the validated utility y.

$$\varphi_c \;\propto\; I(s_c;\, y)\,, \qquad \sum_c \varphi_c = 1 \tag{5.16}$$

If mutual information cannot be computed, the variability of each characteristic score serves as

a surrogate measure of informativeness.

$$\varphi_c \propto \mathrm{Var}(s_c), \qquad \sum_c \varphi_c = 1 \tag{5.17}$$

These importances reflect feature saliency, shaping future inference by weighting key quality dimensions in CORTEX and QUANTA. A detailed illustration for the process of FISSION is provided in Algorithm 3.

Algorithm 3: FISSION: Feature Importance Learning

Input: History $\mathcal{H}$ of runs with characteristic scores $\mathbf{s}^{(t)}$ and utilities or validations $y^{(t)}$.

Output: Importance scores $\{\varphi_c\}$ over characteristics C.

Step 1: Build Feature–Target Dataset
Initialize matrix S and vector y
foreach *entry in* $\mathcal{H}$ **do**
    Append row $\mathbf{s}^{(t)}$ to S
    Append $y^{(t)}$ (validated utility if available, else predicted utility) to y

Step 2: Compute Importances
if *Mutual information estimation is available* **then**
    foreach $c \in C$ **do**
        compute $I(s_c; y)$
    Set $\varphi_c \propto I(s_c; y)$ and normalize so that $\sum_c \varphi_c = 1$
else
    foreach $c \in C$ **do**
        compute $\mathrm{Var}(s_c)$
    Set $\varphi_c \propto \mathrm{Var}(s_c)$ and normalize so that $\sum_c \varphi_c = 1$

return $\{\varphi_c\}$

5.2.1.4 ECHO Implementation

ECHO implements lifelong learning in a structured manner, enabling all agent behaviours, including mappings, weights, and importances, to improve over time and remain synchronized. The ECHO agent serves as the memory module within AURA, ensuring that each execution of

the AURA pipeline appends a record to a persistent learning store that includes dataset context, computed metrics, user targets, thresholds, weights, and utility scores. Each entry in $\mathcal{H}$ records the complete state of a run, thereby providing a comprehensive training signal for future updates and supporting progressive learning by refining QUANTA and CORTEX with accumulated data.

$$\mathcal{H} = (x, s, q_c, \tau_c, w, U(X; x), y_{\text{val}}) \tag{5.18}$$

Here, y_{val} denotes an optional user-provided validation.

If post-hoc validation feedback is provided by the user, such as an approved utility score, the system utilizes this supervision to refine mappings and weights. Following each execution, ECHO updates:

- QUANTA's label mappings,
- CORTEX's regression model via new context–weight pairs, and
- FISSION's importance scores from utility–feature relations.

Formally, the global optimization objective integrates these learning loops and encapsulates AURA's goal: to align predicted utility with validated utility, enforce ordered qualitative mappings, and prevent overfitting.

$$\min_{m,\ \{\theta_c, b_c\}_{c \in C}} \sum_{t=1}^{T} \left[\underbrace{\left\| U(\mathbf{X}_t; x_t) - y_t^{\text{val}} \right\|_2^2}_{\text{utility fit}} + \underbrace{\gamma\, \Phi(m)}_{\text{isotonic/order regularization}} + \underbrace{\lambda \sum_{c \in C} \|\theta_c\|_2^2}_{\text{model regularization}} \right] \tag{5.19}$$

In this formulation, $\Phi(m)$ imposes a penalty for order violations in label mapping.

A detailed illustration for the process of ECHO is provided in Algorithm 4.

Algorithm 4: ECHO: Experience-Consolidating Heuristic Optimizer

Input: Current history $\mathcal{H}$, new run record $(x, \mathbf{s}, \{q_c\}, \{\tau_c\}, \mathbf{w}, U(\mathbf{X}; x), y_{\text{val}})$, current mapping m, parameters $\{\theta_c, b_c\}_c$.

Output: Updated history $\mathcal{H}$, refined mapping m, and parameters $\{\theta_c, b_c\}_c$.

Step 1: Append Run to History
Insert $(x, \mathbf{s}, \{q_c\}, \{\tau_c\}, \mathbf{w}, U(\mathbf{X}; x), y_{\text{val}})$ into $\mathcal{H}$
If $|\mathcal{H}|$ exceeds a maximum size $T_{\max}$, drop oldest entries

Step 2: Form Training Set
From $\mathcal{H}$, extract:

1. Contexts $\{x^{(t)}\}$
2. Weights $\{\mathbf{w}^{(t)}\}$
3. Utilities and validations $\{U(\mathbf{s}^{(t)}; x^{(t)}),\ y_{\text{val}}^{(t)}\}$
4. Qualitative labels $\{q_c^{(t)}\}$ and thresholds $\{\tau_c^{(t)}\}$

Step 3: Optimize Global Objective
Update m and $\{\theta_c, b_c\}_c$ to minimize:

$$\min_{m,\ \{\theta_c, b_c\}_{c \in C}} \sum_{t=1}^{T} \left[\left\| U(\mathbf{X}_t; x_t) - y_t^{\text{val}} \right\|_2^2 \ + \ \gamma\, \Phi(m) \ + \ \lambda \sum_{c \in C} \|\theta_c\|_2^2 \right]$$

Implement via alternating optimization:

1. Fix $\{\theta_c, b_c\}_c$ and update m using QUANTA-like updates with isotonic projection
2. Fix m and refit $\{\theta_c, b_c\}_c$ using CORTEX's ridge regression

return Updated $\mathcal{H}$, m, and $\{\theta_c, b_c\}_c$

5.2.1.5 Integrated AURA Sub-Agent Implementation

The integration of QUANTA, CORTEX, FISSION, and ECHO establishes a self-adaptive ecosystem where all functions co-evolve through supervised learning and feedback mechanisms. AURA can be mathematically represented as follows:

$$U_t = \sum_{c \in C} w_c(x_t;\ \theta_c, b_c)\ s_c(\mathbf{X}_t), \qquad \tau_c^{(t)} = m_t(q_c^{(t)}), \quad \text{with} \tag{5.20}$$

$$(m_{t+1}, \{\boldsymbol{\theta}_c, b_c\}_c) = \arg \min_{m,\, \{\boldsymbol{\theta}_c, b_c\}_{c \in C}} \sum_{t=1}^{T} \left[\left\| U(\mathbf{X}_t; x_t) - y_t^{\text{val}} \right\|_2^2 \; + \; \gamma\, \Phi(m) \; + \; \lambda \sum_{c \in C} \|\boldsymbol{\theta}_c\|_2^2 \right] \tag{5.21}$$

where, t = Execution index (current AURA run),

$\mathbf{X}_t$ = Dataset evaluated at run t,

C = Set of data-quality characteristics,

c = Index of a characteristic, $c \in C$,

$s_c(\mathbf{X}_t)$ = Normalized score of characteristic c for dataset $\mathbf{X}_t$,

x_t = Dataset context vector at run t,

$w_c(x_t; \boldsymbol{\theta}_c, b_c)$ = Context-aware weight for characteristic c inferred by CORTEX,

$q_c^{(t)}$ = User-provided qualitative label for characteristic c,

$m_t(\cdot)$ = QUANTA mapping from qualitative labels to numeric thresholds at run t,

$\tau_c^{(t)}$ = Numeric threshold for characteristic c at run t,

U_t = Overall utility score produced by AURA at run t,

y_t^{val} = User-validated or approved utility score (if available),

$\boldsymbol{\theta}_c, b_c$ = CORTEX parameters mapping context to importance of c,

$\Phi(m)$ = Regularization enforcing ordered qualitative mappings,

γ, λ = Regularization coefficients, and

$\mathcal{H}$ = Implicit history of past runs stored by ECHO.

Each sub-agent contributes interpretability, adaptability, contextual reasoning, or empirical learning. The primary innovation of AURA, as depicted in Figure 5.2 is the integration of these capabilities into a unified and coherent pipeline:

- QUANTA ensures **human-aligned mapping** between qualitative input and quantitative evaluation.
- CORTEX provides **contextual intelligence**, adapting the weighting scheme to dataset structure.

- FISSION introduces **data-driven explanatory power**, revealing which features truly influence utility.
- ECHO enables **continuous self-improvement**, forming the backbone of adaptive behavior.

Collectively, these components form a fully agentic, self-adjusting, and transparent system for data utility assessment. The model adapts to new datasets, maintains interpretability through structured mappings and weights, and bases utility predictions on both mathematical reasoning and empirical evidence. A detailed illustration for the process of AURA is provided in Algorithm 5.

Algorithm 5: AURA: Master Orchestration

Input: Dataset $\mathbf{X}$, user JSON with qualitative targets $\{q_c\}$, optional target utility $U^\star$, optional validation y_{val}.

Output: Overall utility $U(\mathbf{X}; x)$, thresholds $\{\tau_c\}$, weights $\mathbf{w}(x)$, importance scores $\{\varphi_c\}$, threshold satisfaction $\{\text{meet}_c\}$.

Phase 1: Profiling and Context Extraction

Compute characteristic scores $\mathbf{s} = (s_{\text{comp}}, s_{\text{uniq}}, s_{\text{div}}, s_{\text{fmt}})$

Compute context vector x (e.g., rows, columns, % missing, cardinality)

Phase 2: Qualitative → Quantitative via QUANTA

Call QUANTA with $(\{q_c\}, \mathbf{s}, U^\star)$

Obtain thresholds $\{\tau_c\}$ and updated mapping m

Phase 3: Context-Aware Weighting via CORTEX

Call CORTEX with $(x, \mathcal{H})$ to obtain weights $\mathbf{w}(x)$

Phase 4: Utility Computation and Threshold Evaluation

Compute utility $U(\mathbf{X}; x) = \sum_c w_c(x)\, s_c$

foreach $c \in C$ **do**

  $\text{meet}_c \leftarrow \mathbb{I}\{s_c \geqslant \tau_c\}$

Compute MeetRate $= \dfrac{1}{|C|} \sum_c \text{meet}_c$

Phase 5: Feature Importance via FISSION

Call FISSION with $(\mathcal{H})$ to obtain feature importances $\{\varphi_c\}$

Phase 6: History Update and Learning via ECHO

Construct run record: $(x, \mathbf{s}, \{q_c\}, \{\tau_c\}, \mathbf{w}(x), U(\mathbf{X}; x), y_{\text{val}})$

Call ECHO to update history $\mathcal{H}$

Refine $(m, \{\theta_c, b_c\})$ by minimizing the global objective

return $(U(\mathbf{X}; x), \{\tau_c\}, \mathbf{w}(x), \{\varphi_c\}, \{\text{meet}_c\})$

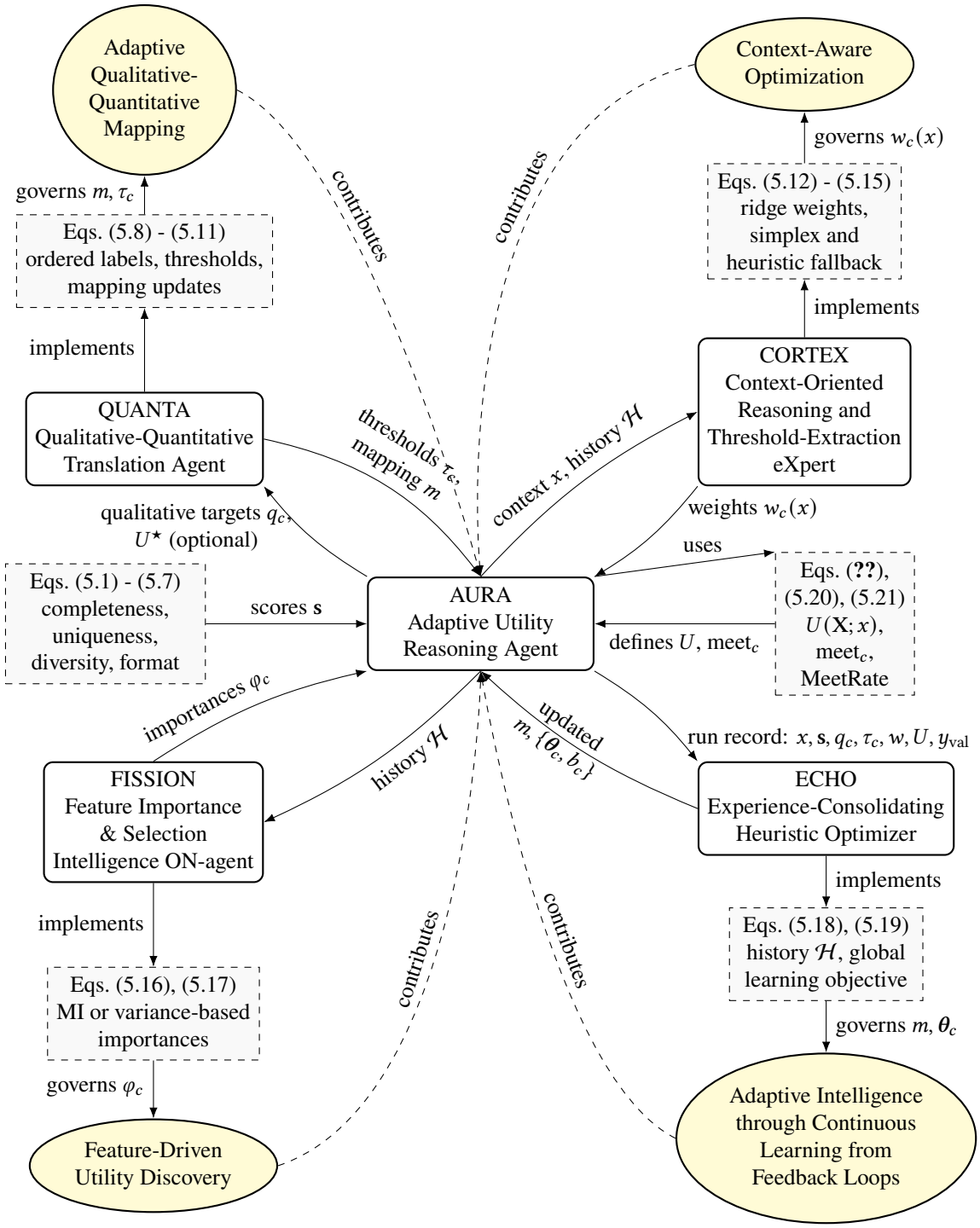


Figure 5.2: Unified architectural view and interaction diagram of AURA, QUANTA, CORTEX, FISSION and ECHO with underlying equations and associated research contributions.

5.2.2 MAESTRO Implementation

MAESTRO uses the user-defined data utility target and the user-provided dataset as inputs to achieve its objectives of data utility and privacy optimization. As depicted previously in Figure 5.1, it comprises multiple sub-agents, each providing multiple functionalities as described in the following sections. At a high level, MAESTRO organizes privacy-utility optimization for textual data as: **SENTINEL** (profiling) → **CHIMERA** (synthetic augmentation) → **AEGIS** (DP mechanism selection + calibration) → **ORION** (utility evaluation) under the governance of **MAESTRO** (orchestration & persistent policy).

Mathematically, let $\mathcal{D} = \{t_i\}_{i=1}^{N}$ be text data, with embeddings $X \in \mathbb{R}^{N\times d}$ (rows x_i). SENTINEL reports sensitivity $Sens \in [0, \infty)$ and semantic density $\rho \in [0, 1]$. CHIMERA produces $\tilde{\mathcal{D}}$ and the augmented $\mathcal{D}' = \mathcal{D} \cup \tilde{\mathcal{D}}^{\star}$ after filtering. AEGIS returns a mechanism $\mathcal{M} \in \{\text{Laplace}, \text{Gaussian}\}$ and $(\epsilon^{\star}, \delta)$; ORION computes utility U from cosine retention and clustering stability.

5.2.2.1 SENTINEL Implementation

SENTINEL offers adaptive intelligence by converting qualitative privacy concerns into quantitative indicators (S and ρ). It further conducts feature-driven utility discovery by identifying specific features, such as columns, tokens, and structures, that influence privacy and utility. Additionally, SENTINEL supports context-aware optimization by profiling each dataset individually, rather than relying on uniform assumptions. The system outputs $\{\mathbf{X}, S, \rho, N, d, \text{columns used}\}$ for downstream agents and to guide the appropriate level of privacy noise to introduce in MAESTRO. SENTINEL quantifies the privacy and utility impact of each record and characterizes the distribution of the data. The SENTINEL agent provides the following functions to determine the sensitivity S and density ρ for each dataset.

- Selects user-specified privacy-sensitive text columns or, if not specified, applies automatic detection of text fields using data-type heuristics, such as identifying common PII fields

including emails and phone numbers.

- Scans text for regular expression-based PII patterns, such as emails, phone numbers, and long numeric sequences, and calculates both the PII hit rate and lexical rarity.
- Constructs embeddings from the text $\mathbf{X}$ using either TF–IDF or a configured embedding function, applied to the concatenated sensitive columns for each row.
- Computes semantic density using the embeddings, defined as the fraction of non-zero entries in the embedding matrix ($\rho = \|\mathbf{X}\|_0/(Nd)$), and calculates sensitivity, which measures the impact of a single row on the vector set based on PII hit rate and word rarity ($S = \mathrm{clip}\big(0.1 + 0.3 \cdot \mathrm{PIIrate} + 1.2 \cdot \mathrm{Rarity},\ 0.1,\ 2.0\big)$).

$$X = \mathrm{Embed}(\mathcal{D}), \qquad \rho = \frac{\|X\|_0}{Nd} \tag{5.22}$$

Sensitivity estimation (clipped) uses PII hits per 1k tokens and hapax fraction π:

$$Sens = \mathrm{clip}\left(0.1 + a \cdot \frac{\sum_i h(t_i)}{\sum_i |t_i|/1000} + b \cdot \pi,\ 0.1,\ 2.0\right) \tag{5.23}$$

A detailed illustration for the process of SENTINEL is provided in Algorithm 6.

Algorithm 6: SENTINEL: Profiling and Sensitivity Estimation

Input: Dataset path d, global sensitive columns $\mathcal{S}_{\text{global}}$, per-file sensitive columns $\mathcal{S}_{\text{file}}$.

Output: Profile struct with sensitivity S, semantic density ρ, selected columns, and basic stats; list of concatenated texts per row.

Step 1: Load and select sensitive columns

Load CSV into dataframe df

if $d \in \mathcal{S}_{\text{file}}$ **then**

  cols $\leftarrow \mathcal{S}_{\text{file}}[d] \cap$ df.columns

else

  if $\mathcal{S}_{\text{global}}$ *not empty* **then**

    cols $\leftarrow \mathcal{S}_{\text{global}} \cap$ df.columns

  else

    cols $\leftarrow$ AutoDetectTextColumns(df)

If cols is empty, fall back to first text-like column (if any)

Step 2: Build row-wise text and estimate sensitivity

Construct Texts = $\{t_i\}$ by concatenating the values of cols for each row

Initialize PII hit count and vocabulary statistics

foreach $t_i \in$ *Texts* **do**

  Detect PII patterns (e-mail, phone, IDs) and accumulate hits

  Tokenize t_i and update token counts

Compute PIIrate = PII hits per 1000 tokens

Compute Rarity = fraction of tokens with count 1

Set

$$S \leftarrow \text{clip}\big(0.1 + 0.3\,\text{PIIrate} + 1.2\,\text{Rarity},\ 0.1,\ 2.0\big)$$

Step 3: Compute embeddings and density

Compute embeddings $\mathbf{X} \leftarrow$ Embed(Texts) $\in \mathbb{R}^{N \times d}$

Compute $\rho \leftarrow \|\mathbf{X}\|_0/(Nd)$ (semantic density)

Step 4: Build profile

profile $\leftarrow \{S, \rho, \text{cols}, N, d\}$

return (profile, Texts)

5.2.2.2 CHIMERA Implementation

CHIMERA incorporates adaptive intelligence by activating only when utility cannot be achieved with existing data. It applies context-aware optimization through dataset-specific augmentation ratios and supports feature-driven utility discovery by reinforcing features pertinent to utility. CHIMERA

generates synthetic text to improve the semantic structure of data before noise is introduced. When utility requirements are unmet under strict privacy budgets, CHIMERA augments the dataset with additional examples to stabilize cluster structure and enhance noise tolerance. If SENTINEL's output signals data sparsity or requires a lower ε, CHIMERA generates synthetic data at a controlled ratio α (from 0% to 40%), thereby extending the original dataset. The CHIMERA agent performs the following functions to generate the augmented corpus $\mathcal{D}' = \mathcal{D} \cup \tilde{\mathcal{D}}^{\star}$.

- Constructs topic-conditioned text sequences using a large language model (LLM) or, if unavailable, employs an n-gram model as a fallback.
- Adjusts the augmentation ratio $\alpha \in \{0, 0.1, 0.25, 0.4\}$ to maintain controlled semantic drift.
- Filters low-quality or duplicate synthetic data before merging with the corpus to construct $\mathcal{D}'$.

In summary, CHIMERA generates augmented texts by embedding synthetic data $\mathbf{X}_{\alpha}$, which SENTINEL recomputes when $\alpha > 0$.

For $\alpha \in \mathcal{A}$, the system synthesizes $\tilde{\mathcal{D}}$ using the generator $\mathcal{G}_{\theta}$ (either an LLM or n-gram model), and filters the results using predicate Q:

$$\tilde{\mathcal{D}} = \mathcal{G}_{\theta}(\mathcal{D}, \alpha), \quad \tilde{\mathcal{D}}^{\star} = \{\tilde{t} \in \tilde{\mathcal{D}} \mid Q(\tilde{t}) = 1\}, \quad \mathcal{D}' = \mathcal{D} \cup \tilde{\mathcal{D}}^{\star} \tag{5.24}$$

The process iterates over $\alpha \in \{0, 0.1, 0.25, 0.4\}$ to achieve the target utility at a lower ε. By incorporating realistic synthetic data, CHIMERA reduces the need for noise addition, enabling a smaller ε and minimizing adverse effects on data utility. A detailed illustration for the process of CHIMERA is provided in Algorithm 7.

Algorithm 7: CHIMERA: Synthetic Text Augmentation

Input: Original texts Texts = $\{t_i\}_{i=1}^{N}$, augmentation ratio $\alpha \in [0, 1]$.
Output: Augmented text collection Texts$_\alpha$.

Step 1: Determine number of new samples
$n_{\text{new}} \leftarrow \max(1, \lfloor \alpha N \rfloor)$;

Step 2: Generate candidate synthetic texts
if *LLM or fine-tuned model available* **then**
    Candidates $\leftarrow$ LLM_GenerateInDomain(Texts, n_{new});
else
    Candidates $\leftarrow$ NGram_FallbackGenerate(Texts, n_{new});

Step 3: Quality filtering
Accepted $\leftarrow$ [ ], Seen $\leftarrow \emptyset$;
foreach $t \in$ *Candidates* **do**
    if *Length*$(t) < L_{\min}$ **then**
        continue;
    key $\leftarrow$ NormalizeKey(t);
    if *key* $\in$ *Seen* **then**
        continue;
    if *QualityGate_IsPlausible*(t) = false **then**
        continue;
    Append t to Accepted;
    Seen $\leftarrow$ Seen $\cup$ {key};

Step 4: Merge with original texts
Texts$_\alpha \leftarrow$ Texts $\cup$ Accepted;

return Texts$_\alpha$;

5.2.2.3 AEGIS Implementation

AEGIS performs dynamic privacy-utility optimization by adjusting the noise scale to maintain utility. This process is context-aware, leveraging sensitivity S and density ρ to adapt the ε-grid. AEGIS generates a set of candidate ε values (for example, 0.1, 0.2, and so on) and uses the previously computed S and ρ to determine the required noise for each mechanism, specifically Laplace and Gaussian. After determining the appropriate noise, AEGIS applies it to the embeddings and transmits both noisy versions to ORION. The system then selects the smallest ε and mechanism

$\mathcal{M} \in \{\text{Laplace}, \text{Gaussian}\}$ such that $U(\mathbf{X}') \geqslant U_{\text{target}}$.

- Generates an adaptive ε-grid, optionally informed by prior runs, and sets δ as a fixed parameter.
- Evaluates both Laplace and Gaussian mechanisms to minimize utility loss.
- Computes L1 and L2 sensitivities and dynamically derives b and σ.

L1 and L2 sensitivities are derived from S and ρ:

$$L1 = S\left(1 + \tfrac{1}{2}\rho\right) \tag{5.25}$$

$$L2 = S\sqrt{1 + \rho} \tag{5.26}$$

For each ε in the grid, AEGIS generates privatized embeddings using the following mechanisms. Laplace mechanism

$$b = \frac{L1}{\varepsilon}, \qquad \mathbf{X}'_{\text{Lap}} = \mathbf{X} + \text{Lap}(0, b) \tag{5.27}$$

and the Gaussian mechanism

$$\sigma = \frac{c\, L2\sqrt{2\ln(1.25/\delta)}}{\varepsilon}, \qquad \mathbf{X}'_{\text{Gau}} = \mathbf{X} + \mathcal{N}(0, \sigma^2\mathbf{I}) \tag{5.28}$$

Both candidates are then forwarded to ORION for utility evaluation. In summary, AEGIS translates the theoretical privacy budget into practical noisy data.

The overall optimization objective function for AEGIS is denoted as:

$$(\mathcal{M}^*, \varepsilon^*) = \arg\min_{\mathcal{M},\varepsilon} \varepsilon \quad \text{s.t. } U(\mathbf{X}') \geqslant U_{\text{target}} \tag{5.29}$$

A detailed illustration for the process of AEGIS is provided in Algorithm 8.

Algorithm 8: AEGIS: Sensitivity-Aware DP Mechanism Selection

Input: Embeddings $\mathbf{X}$, sensitivity S, semantic density ρ, target utility U_{target}, prior preference $\varepsilon_{\text{pref}}$ (optional).

Output: Chosen configuration sel = $(\mathcal{M}^*, \varepsilon^*, \delta^*, U^*)$, full grid of evaluated candidates.

Step 1: Build ε-grid and δ

Construct base grid $\mathcal{E}_{\text{base}}$ (e.g., $\{0.2, 0.3, 0.4, \ldots, 4.0\}$);
Adjust grid using S and $\varepsilon_{\text{pref}}$ (scaling/compression);
Set δ to a small constant (e.g., 10^{-5});
$\mathcal{E} \leftarrow$ sorted, de-duplicated union of scaled values;

Step 2: Enumerate Laplace and Gaussian candidates

Results $\leftarrow$ [ ];
foreach $\varepsilon \in \mathcal{E}$ **do**
    Compute $\Delta_1 = S(1 + \frac{1}{2}\rho)$ and $\Delta_2 = S\sqrt{1+\rho}$;
    Laplace candidate
    $b \leftarrow \Delta_1/\varepsilon$;
    $\mathbf{X}'_{\text{Lap}} \leftarrow \mathbf{X} + \textsc{LaplaceNoise}(0, b)$;
    $(U_{\text{Lap}}, m_{\text{Lap}}) \leftarrow \text{ORION\_}\textsc{Utility}(\mathbf{X}, \mathbf{X}'_{\text{Lap}})$;
    Append $(\text{Laplace}, \varepsilon, 0, U_{\text{Lap}}, m_{\text{Lap}})$ to Results;
    Gaussian candidate
    $\sigma \leftarrow c\,\Delta_2\sqrt{2\ln(1.25/\delta)}/\varepsilon$;
    $\mathbf{X}'_{\text{Gau}} \leftarrow \mathbf{X} + \textsc{GaussianNoise}(0, \sigma)$;
    $(U_{\text{Gau}}, m_{\text{Gau}}) \leftarrow \text{ORION\_}\textsc{Utility}(\mathbf{X}, \mathbf{X}'_{\text{Gau}})$;
    Append $(\text{Gaussian}, \varepsilon, \delta, U_{\text{Gau}}, m_{\text{Gau}})$ to Results;

Step 3: Choose smallest ε meeting U_{target}

$\mathcal{F} \leftarrow \{r \in \text{Results} : r.U \geqslant U_{\text{target}}\}$;
if $\mathcal{F} \neq \emptyset$ **then**
    Sort $\mathcal{F}$ by increasing ε, breaking ties by decreasing U;
    sel $\leftarrow$ first element of $\mathcal{F}$;
    sel.meetsTarget $\leftarrow$ true;
else
    sel $\leftarrow \arg\max_{r \in \text{Results}} r.U$;
    sel.meetsTarget $\leftarrow$ false;

return (sel, Results);

5.2.2.4 ORION Implementation

ORION facilitates feature-driven utility assessment by determining which embedding features are essential for maintaining semantic content. In alignment with AEGIS, it provides context-aware optimization by addressing the structural characteristics of each dataset individually. ORION measures the preservation of meaning, similarity, and structural integrity in data following the introduction of noise by AEGIS. Cosine similarity is used to evaluate the correspondence between original and noised embeddings, and silhouette scores assess the stability of clustering structures. For every noisy dataset generated by AEGIS, ORION implements the following procedures to evaluate the utility of the perturbed data for downstream tasks such as clustering or classification:

- Measures embedding cosine retention, which quantifies the extent to which each vector preserves its direction following the introduction of noise.

$$u_{\cos} = \frac{1}{M} \sum_{i=1}^{M} \frac{\mathbf{x}_i^\top \mathbf{x}_i'}{\|\mathbf{x}_i\| \, \|\mathbf{x}_i'\|} \tag{5.30}$$

- Assesses deviation in cluster structure using the silhouette score, specifically evaluating whether Δsil indicates substantial drift.

$$\Delta\text{sil} = \text{sil}(\mathbf{X}') - \text{sil}(\mathbf{X}) \tag{5.31}$$

- Combines these metrics using weighted aggregation to compute an overall utility U. If $U \geqslant U_{\text{target}}$, ORION deems the results sufficient.

$$U = w_1 u_{\cos} + w_2 \max(-\gamma, \Delta\text{sil}) \tag{5.32}$$

$$\mathcal{F} = \left\{ \left(\varepsilon, U(\varepsilon, \mathcal{M})\right) : \ \varepsilon \in \mathcal{E}, \ \mathcal{M} \in \{\text{Lap,Gau}\} \right\} \tag{5.33}$$

If no configuration satisfies $U \geqslant U_{\text{target}}$ at α, ORION selects $\arg\max U$ and reports the shortfall $U - U_{\text{target}}$.

A detailed illustration for the process of ORION is provided in Algorithm 9.

Algorithm 9: ORION: Utility Computation and Assessment

Input: Original embeddings $\mathbf{X}$, DP configuration sel = $(\mathcal{M}^*, \varepsilon^*, \delta^*)$ (from AEGIS).

Output: Utility metrics (cosine retention, silhouette scores, aggregate utility U).

Step 1: Reconstruct noisy embeddings under chosen mechanism
if $\mathcal{M}^* = \mathsf{Laplace}$ **then**
  Recompute Δ_1 and scale $b = \Delta_1/\varepsilon^*$ (using S, ρ);
  $\mathbf{X}' \leftarrow \mathbf{X} + \textsc{LaplaceNoise}(0, b)$;
else
  Recompute Δ_2 and $\sigma = c\,\Delta_2\sqrt{2\ln(1.25/\delta^*)}/\varepsilon^*$;
  $\mathbf{X}' \leftarrow \mathbf{X} + \textsc{GaussianNoise}(0, \sigma)$;

Step 2: Optional subsampling for scalability
if $N = \mathit{rows}(\mathbf{X})$ *is large* **then**
  Randomly select index subset $\mathcal{I}$ of size M and define $\mathbf{A} = \mathbf{X}_{\mathcal{I}}$, $\mathbf{B} = \mathbf{X}'_{\mathcal{I}}$;
else
  $\mathbf{A} \leftarrow \mathbf{X}$; $\mathbf{B} \leftarrow \mathbf{X}'$;

Step 3: Cosine retention
Compute $u_{\cos} = \frac{1}{M}\sum_{i=1}^{M} \frac{\mathbf{a}_i^\top \mathbf{b}_i}{\|\mathbf{a}_i\|\|\mathbf{b}_i\|}$;

Step 4: Cluster structure and Δsil
Choose k (e.g., based on $\sqrt{M}$);
Cluster $\mathbf{A}$ and $\mathbf{B}$ using k-means to obtain labels;
Compute sil($\mathbf{A}$) and sil($\mathbf{B}$), then $\Delta\text{sil} = \text{sil}(\mathbf{B}) - \text{sil}(\mathbf{A})$;

Step 5: Aggregate utility
Compute $U = w_1 u_{\cos} + w_2 \max(-\gamma, \Delta\text{sil})$;

return $\{u_{\cos}, \Delta\text{sil}, U, \text{sil}(\mathbf{A}), \text{sil}(\mathbf{B}), \mathcal{M}^*, \varepsilon^*, \delta^*\}$;

5.2.2.5 Integrated MAESTRO Sub-Agent Implementation

MAESTRO innovation incorporates adaptive intelligence supported by continual learning through a policy cache, context-aware optimization, and feature-driven utility discovery to achieve dynamic data privacy and utility optimization. As depicted in Figure 5.3, MAESTRO coordinates

the four sub-agents in a feedback loop that:

1. profiles the dataset (SENTINEL),

2. activates CHIMERA augmentation only if needed,

3. attempts DP calibration (AEGIS) evaluated by ORION,

4. and learns from past runs to accelerate future optimization.

While executing all the sub-agents in order, it ensures reproducibility by writing a JSON report and updating a learning cache with the best ε for each type of tabular data. This approach helps future executions to leverage the learnings and save time during optimization.

$$\epsilon_{\text{pref}} \leftarrow f\big(\varepsilon^{\star}, U, \text{feedback}\big), \qquad \mathcal{E} \leftarrow \text{adj}(\mathcal{E}, \varepsilon). \tag{5.34}$$

A detailed illustration for the process of MAESTRO is provided in Algorithm 10.

Algorithm 10: MAESTRO: Target-Aware Orchestration

Input: Dataset list $\mathcal{D}$, global sensitive columns $\mathcal{S}_{\text{global}}$, per-file sensitive columns map $\mathcal{S}_{\text{file}}$, target utility U_{target}, augmentation grid $\mathcal{A} = \{\alpha_1, \ldots, \alpha_L\}$, PolicyCache (optional prior $\varepsilon_{\text{pref}}$ per dataset).

Output: Per-dataset reports with chosen $(\mathcal{M}^\star, \varepsilon^\star, \delta^\star, \alpha^\star)$ and recorded utilities.

Step 1: Loop over datasets
foreach $d \in \mathcal{D}$ **do**
  (profile, Texts) ← SENTINEL_Profile($d, \mathcal{S}_{\text{global}}, \mathcal{S}_{\text{file}}$);
  Initialize bestSel ← ∅, bestReport ← ∅, metTarget ← false;
  $\varepsilon_{\text{pref}}$ ← PolicyCache$[d].\varepsilon_{\text{pref}}$ (if available);

  Step 2: α-loop with optional augmentation
  foreach $\alpha \in \mathcal{A}$ **do**
    if $\alpha = 0$ **then**
      Texts$_\alpha$ ← Texts;
    else
      Texts$_\alpha$ ← CHIMERA_Augment(Texts, α);
    Compute embeddings X ← Embed(Texts$_\alpha$);

    Step 3: AEGIS selection and ORION evaluation
    (sel, grid) ← AEGIS_SelectForTarget(X, profile.S, profile.ρ, U_{target}, $\varepsilon_{\text{pref}}$);
    metrics ← ORION_Assess(X, sel);
    report ← BuildReport(d, profile, sel, grid, metrics, U_{target}, α);

    Step 4: Update best selection and early stop if target met
    if *sel.meetsTarget* = true **then**
      bestSel ← sel, bestReport ← report, metTarget ← true;
      break from α-loop;
    else
      if *bestSel* = ∅ ***or*** *sel.util* > *bestSel.util* **then**
        bestSel ← sel, bestReport ← report;

  Step 5: Persist learning to PolicyCache and output
  Update PolicyCache$[d]$ with profile.S, profile.ρ, bestSel.ε, bestSel.$\mathcal{M}$, used α, and metTarget;
  Write bestReport to disk or logging sink;

return all per-dataset reports and updated PolicyCache;

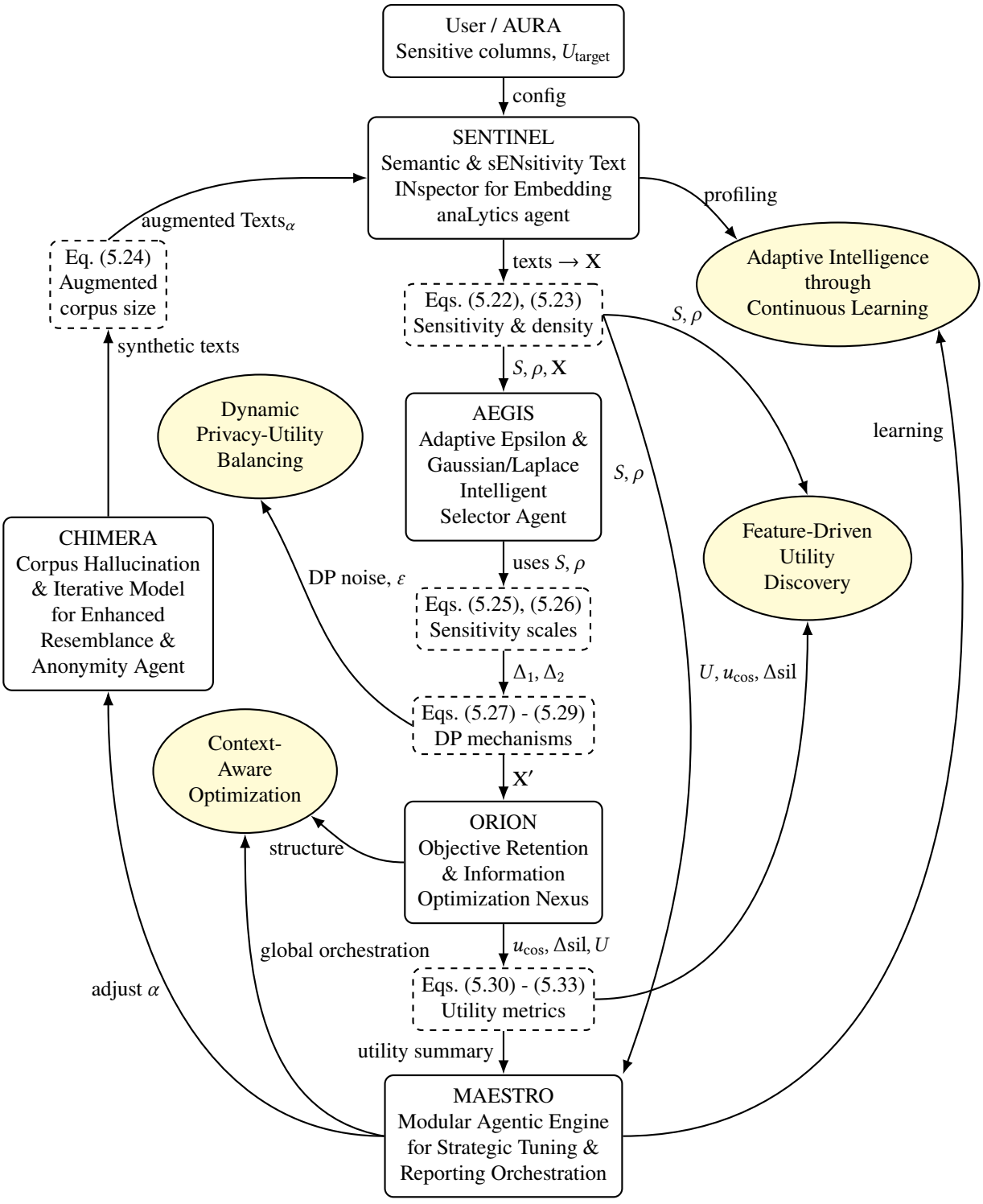


Figure 5.3: Unified architectural view and interaction diagram of MAESTRO, SENTINEL, CHIMERA, AEGIS, and ORION with underlying equations and associated research contributions.

5.2.3 Overall process workflow with combined AURA & MAESTRO behavior

Figures 5.4 and 5.5 present the complete process workflows for AURA and MAESTRO. The AURA workflow initiates by loading the dataset and user preference files, followed by profiling the data through the computation of normalized scores for key quality characteristics such as completeness, uniqueness, diversity, and format consistency. Subsequently, it extracts dataset-level contextual features and determines quantitative thresholds and characteristic weights through the coordinated operation of QUANTA and CORTEX. With these parameters, AURA calculates an overall utility score as a weighted aggregate of the individual characteristics and assesses each characteristic against its respective threshold to determine compliance. All results and outcomes are then stored in a learning repository, which allows the framework to iteratively adapt its mappings, thresholds, and weighting strategies over time.

The MAESTRO workflow builds upon this process by integrating privacy-aware optimization. Provided with dataset paths, identified privacy-sensitive columns, and a user-defined target utility, SENTINEL profiles the data by computing embeddings, sensitivity, semantic density, and related structural attributes. CHIMERA subsequently generates controlled in-domain synthetic text to facilitate privacy preservation, while AEGIS constructs candidate differentially private versions of the dataset across an ε grid and submits them to ORION for evaluation. ORION measures aggregated utility and reporting metrics for each candidate. MAESTRO then selects the configuration that meets the utility target with the minimal privacy budget. If no candidate achieves the target, the system incrementally increases the augmentation factor and repeats the process until the target is met or all options are exhausted. Collectively, AURA and MAESTRO provide an end-to-end framework that translates qualitative user expectations into quantitative utility objectives, adaptively calibrates privacy noise, and retains optimal configurations for future executions.

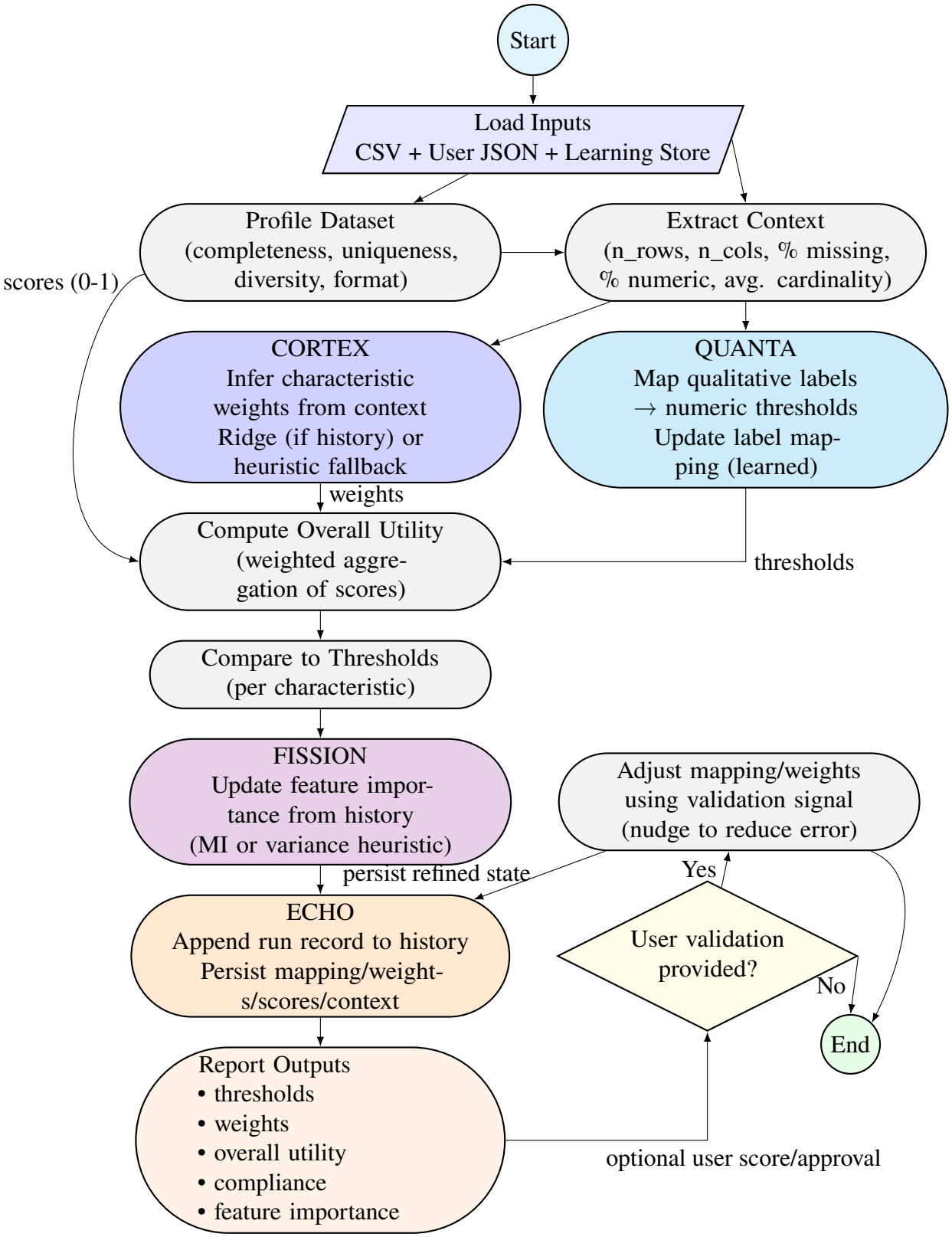


Figure 5.4: AURA Workflow and Sub-Agent Behavior.

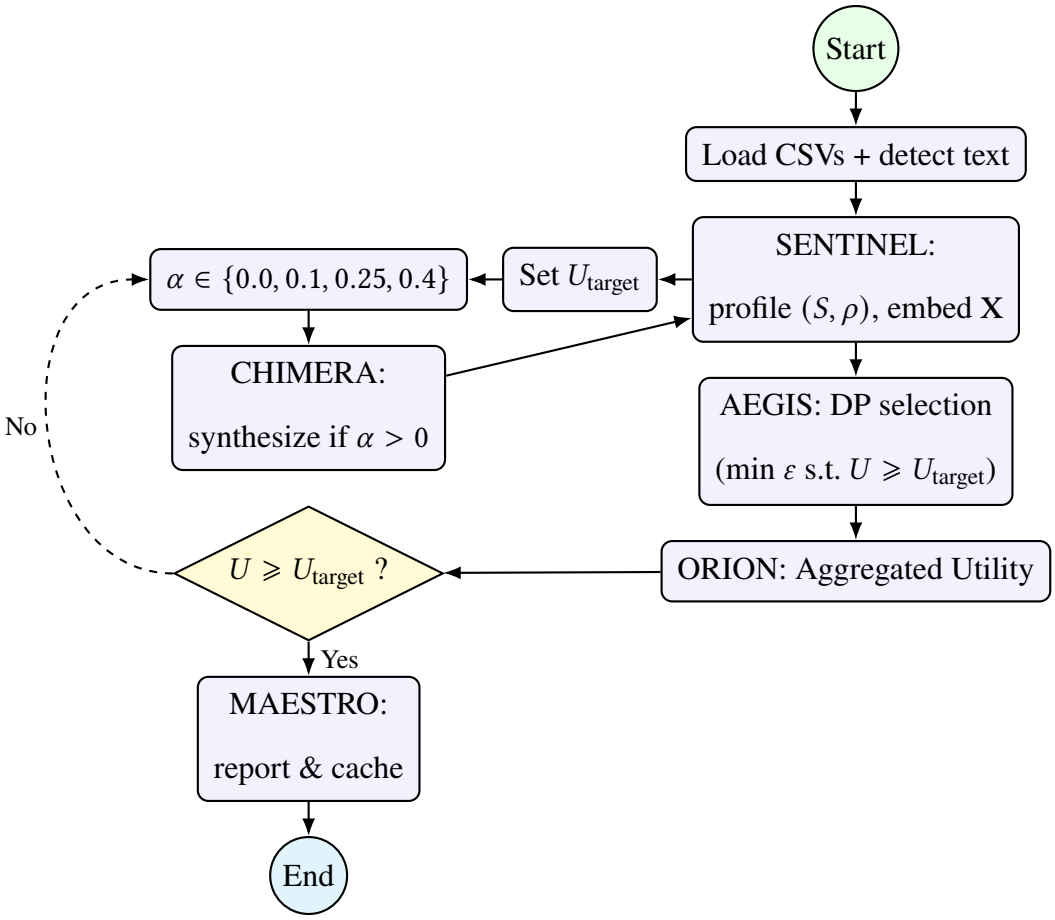


Figure 5.5: MAESTRO Workflow and Sub-Agent Behavior.

5.3 Simulation & Results

To assess the generalizability of the proposed agentic data utility framework, three structurally and semantically distinct datasets were utilized. The first dataset, *BDD100K* [234], is a diverse driving dataset from the University of California, Berkeley. The analysis focused on the attributes *weather*, *time of day*, and *scene*, which are present across 69,863 records. For each record, demographic data points, *age*, *gender*, and *race*, were synthetically generated and appended to facilitate privacy impact analysis; this augmented dataset is referred to as *EV Driving*. During synthetic data generation, the *age* column was distributed as 20%-60%-20% for child, adult, and senior, respectively; *gender* was

balanced at 50% male and 50% female; and *race* was set at 50% white, 20% black, 20% Asian, and 10% other. The final dataset comprised 69,863 fully populated rows, achieving a 100% completeness score. The records were predominantly unique, with only 0.03% of rows duplicated, resulting in a uniqueness fraction of 97%. However, categorical diversity was extremely limited: each attribute (*weather*, *scene*, *time of day*, *race*, *target*) exhibited less than 0.01% distinct values per row, and the demographic fields *age* and *gender* were constant across all entries. Despite this lack of variation, the dataset format was entirely consistent, with every column demonstrating 100% confidence in data type uniformity.

A second dataset, consisting of 3,510,433 time-series records of household energy consumption collected from multiple smart energy meters in the City of London [240], was also analyzed. To examine the effects of dataset size, two subsets were created: one with 1 million records (*Smart Meter-1M*) and another non-overlapping subset with 100,000 records (*Smart Meter-100K*). The 1 million-row dataset exhibited minimal data loss (99.96% completeness), and all rows were unique, resulting in a uniqueness fraction of 100%. The diversity profile indicated that the identifier *LCLid* and the date field *day* had nearly constant values (0.16% and 0.08% distinct values, respectively). In contrast, energy statistics showed moderate variability: *energy_mean* and *energy_sum* varied across approximately 21% of rows, *energy_std* was highly diverse (97.48%), while *energy_median*, *energy_max*, and *energy_min* were largely uniform (less than 1%). The count column was entirely constant (0% diversity). Format consistency was perfect, with strings for *LCLid*, dates for *day*, and floats for all energy measures, each with 100% confidence. The smaller dataset (100,000 records) also achieved 99.96% completeness and 100% uniqueness. Its diversity profile was notably richer: the *day* field varied across 0.83% of rows, and energy statistics such as *energy_mean* and *energy_sum* spanned more than half the dataset (53.13% and 55.88%, respectively). *Energy_median*, *energy_max*, and *energy_min* exhibited 2 to 4% diversity, compared to less than 1% in the larger subset. The count column remained essentially constant (0.03%). Format consistency was again perfect. As a result, models trained on the 100,000-row dataset are likely to encounter greater

temporal and energy pattern diversity, whereas the larger dataset may require more robust handling of imbalanced or uniform features. The consistent data format in both subsets allows pre-processing to focus on diversity differences rather than type mismatches.

The third dataset, referred to as *EV Charging*, consists of 44,225 records of municipal-level EV charging data for the fourth quarter of 2021 [241]. Only 9.20% of the cells are populated, indicating a highly sparse table with numerous missing entries. Record uniqueness is low, with only 4.08% of rows being distinct, suggesting that most observations are repeated or incomplete. Column diversity is similarly limited; the *charging event* field exhibits the highest variation at 4.08%, while categorical descriptors such as *Borough* and *Operator* are nearly constant (0.01% and 0%, respectively). All columns demonstrate perfect type consistency, except for *Total kWh*, which is a float in 95.2% of its entries. The remaining 4.8% of this column likely contains non-numeric or missing values.

5.3.1 AURA implementation across the datasets

AURA was evaluated using four heterogeneous datasets, each characterized by varying data scales, structures, and levels of completeness. As shown in Table 5.3, the *EV Driving* dataset demonstrated moderate utility but was penalized for low uniqueness resulting from duplicate scenes. In contrast, both the *Smart Meter-1M* and *Smart Meter-100K* datasets achieved the highest overall utility, attributed to perfect uniqueness and format scores. The *EV Charging* dataset, despite experiencing extreme incompleteness with approximately 96% missing data, maintained moderate utility due to its diversity and format consistency.

QUANTA's learned mappings from "very low" to "very high" exhibited minor variation across datasets, demonstrating adaptive calibration in response to user-specified overall utility targets (Figure 5.6(a)) and observed performance. As shown in Table 5.4, the "high" category adjusted between 0.71 and 0.76, indicating that QUANTA refines numeric thresholds based on contextual feedback. In the case of the *EV Charging* dataset, which was characterized by low data quality, QUANTA's "high" threshold decreased slightly (0.713), thereby avoiding over-penalization and

Table 5.3: Quantitative Summary of AURA Evaluation.

Dataset	Records	Completeness	Uniqueness	Diversity	Format	Overall Utility	Meets Thresholds (%)
EV Driving	69,863	1.000	0.039	0.804	1.000	0.706	75
Smart Meter-100K	100,000	0.9996	1.000	0.826	1.000	0.864	100
Smart Meter-1M	1,000,000	0.9996	1.000	0.796	1.000	0.796	100
EV Charging	44,225	0.041	0.041	0.761	1.000	0.404	50

illustrating effective self-calibration. The mapping consistently preserved ordinal integrity (very low < low < moderate < high < very high) as a result of the agent's constraint enforcement.

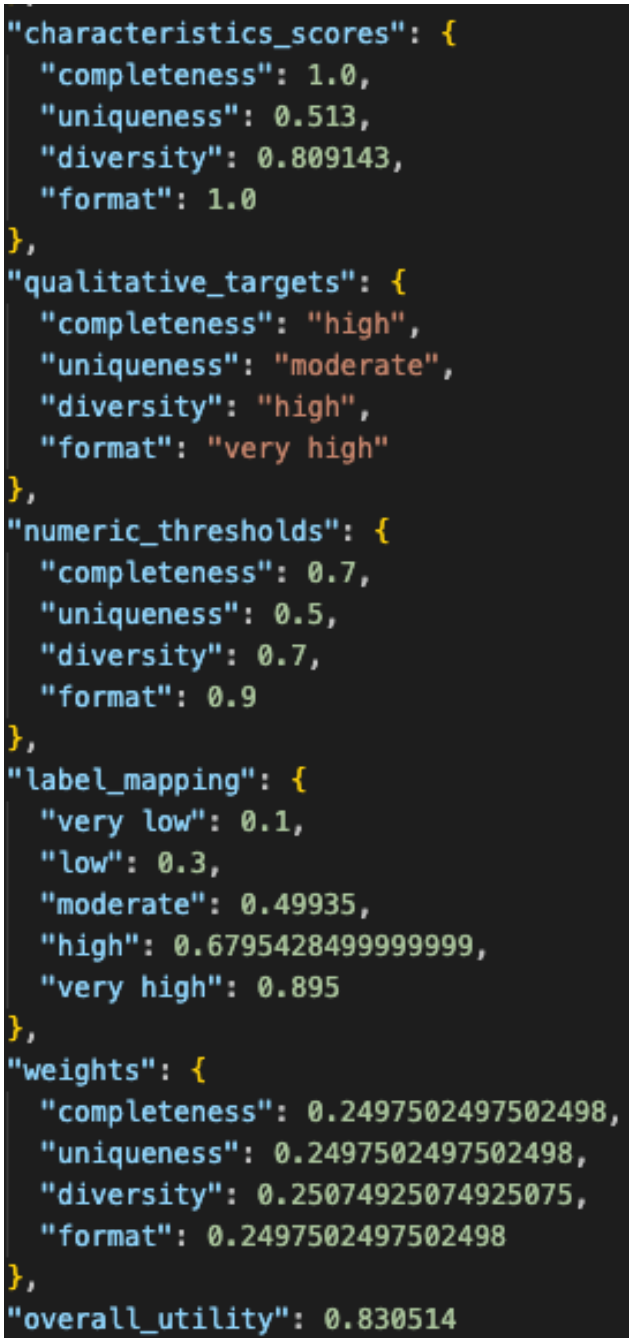
Table 5.4: QUANTA: Learned Label-to-Numeric Mapping per Dataset.

Dataset	Very Low	Low	Moderate	High	Very High
EV Driving	0.10	0.456	0.499	0.733	0.895
Smart Meter-100K	0.10	0.436	0.499	0.759	0.895
Smart Meter-1M	0.10	0.389	0.499	0.754	0.895
EV Charging	0.10	0.435	0.499	0.713	0.895

Across datasets, QUANTA refined its label-to-threshold mapping so that the "high" threshold stabilized between 0.71 and 0.76, indicating internal consistency in AURA's interpretation of user expectations. For instance, in the *EV Charging* dataset, QUANTA lowered the "high" threshold to 0.713 due to extensive missing data (96%), thereby reducing evaluation rigor. In contrast, for the *Smart Meter-100K* dataset, it maintained a stricter "high" threshold at 0.759, reflecting user satisfaction with near-complete data. This adaptive calibration illustrates the learning-based flexibility of AURA's interpretive layer, aligning subjective human intent with objective dataset performance.

As shown in Table 5.5, for the *EV Driving* dataset, CORTEX assigned comparable weights

(a) Example User Input Targets for Data Dimensions and Data Utility



(b) Example Results Stored as JSON in the Learning Store

Figure 5.6: Qualitative User Inputs & Persisted Results by AURA (ECHO)

to completeness and format, while reducing emphasis on diversity, which reflects the dataset's structured and consistent labeling. For the *Smart Meter-100K* and *Smart Meter-1M* datasets, high cardinality resulted in diversity dominance (0.78–1.0), consistent with their temporal breadth and feature richness. For the *EV Charging* dataset, which contains 96% missing values, CORTEX prioritized completeness, confirming context-sensitive weighting. The results in Table 5.5 and Figure 5.7 empirically validate that CORTEX mathematically encodes domain adaptability, such as w_{comp} being proportional to missingness and w_{div} being proportional to cardinality.

Table 5.5: CORTEX: Context-Aware Weighting Distribution.

Dataset	**Completeness**	**Uniqueness**	**Diversity**	**Format**	**Dominant Weight**
EV Driving	0.361	0.297	0.045	0.297	Completeness
Smart Meter-100K	0.074	0.072	0.782	0.072	Diversity
Smart Meter-1M	0.000	0.000	1.000	0.000	Diversity
EV Charging	0.366	0.194	0.246	0.194	Completeness

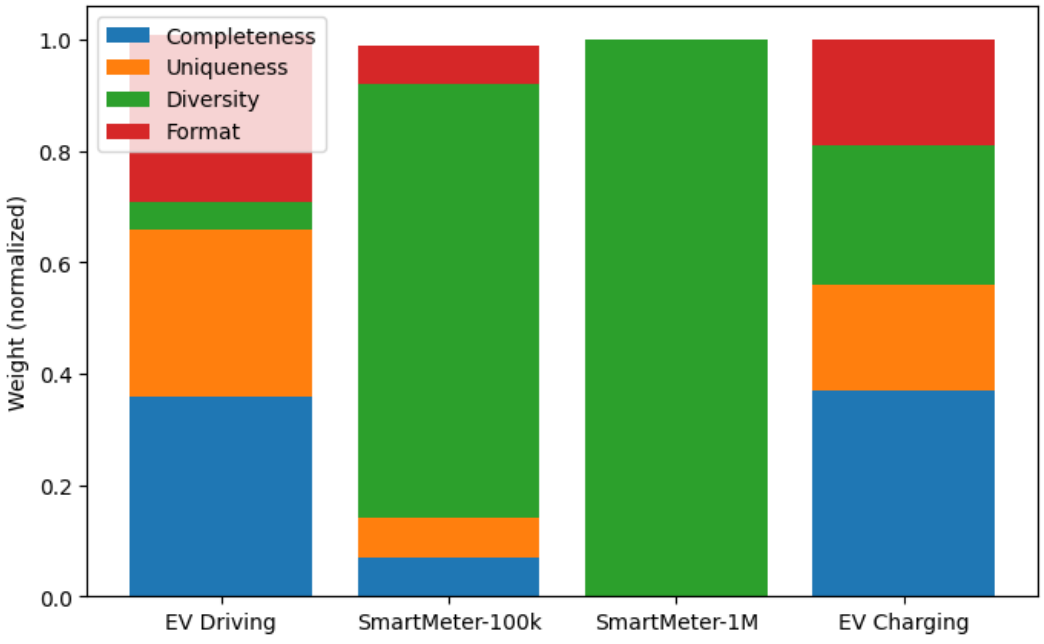


Figure 5.7: CORTEX: Context-Aware Weight Distribution per Dataset.

Table 5.6 shows that, across all analyzed datasets, the format feature consistently exhibited negligible importance ($<10^{-15}$), indicating stable schema conformance. Diversity was the most influential feature in three datasets, whereas uniqueness was dominant in the *Smart Meter-1M* dataset, consistent with its 1 million distinct temporal entries. FISSION's outputs illustrate context-adaptive learning by dynamically adjusting feature priorities in response to changes in the data environment. These results are consistent with theoretical expectations that FISSION emphasizes the characteristic contributing most to perceived utility, thereby demonstrating agent-driven feature selection.

Table 5.6: FISSION: Feature Importance across Datasets.

Dataset	**Completeness**	**Uniqueness**	**Diversity**	**Format**	**Key Driver**
EV Driving	0.139	0.365	0.496	~0.000	Diversity
Smart Meter-100K	0.120	0.428	0.452	~0.000	Diversity & Uniqueness
Smart Meter-1M	0.191	0.505	0.304	~0.000	Uniqueness
EV Charging	0.096	0.374	0.529	~0.000	Diversity

The persistent store in ECHO recorded label mappings, weights, and feature importance over time as depicted in Figure 5.6(b). The operation of ECHO was demonstrated by the consistent evolution of mappings and weights across multiple runs. Analysis of these successive runs revealed the following:

- Stability of mappings across runs, which indicated convergence. Early runs emphasized uniform weighting, whereas later runs demonstrated context-specific adjustments.
- Progressive improvement in weight assignment was observed. For example, the diversity weighting for the *Smart Meter-100K* and *Smart Meter-1M* datasets increased from 0.05 in initial runs to 0.78–1.00 after context-based learning.

- Automatic balance correction occurred when a dataset was highly incomplete. For instance, in the *EV Charging* dataset, ECHO's feedback loop prompted CORTEX to prioritize completeness, demonstrating contextual self-correction.

ECHO incorporates reinforcement via user validation, serving as the cognitive foundation for AURA's adaptive reasoning.

Table 5.7: Analysis of AURA's Overall Capabilities across Datasets.

Capability	**Agent(s)**	**Evidence from Results**	**Demonstrated Behavior**
Adaptive Qualitative-Quantitative Mapping	QUANTA	Dynamic numeric shifts for "high" thresholds (0.713-0.759) across datasets	Adjusted sensitivity to user context and dataset difficulty, maintaining ordinal label order
Context-Aware Categorization	CORTEX	Variation in dominant weights: diversity (*SmartMeter-100K* and *SmartMeter-1M* datasets), completeness (*EV Charging*)	Learned domain relevance through dataset-specific context analysis
Agent-Driven Feature Selection	FISSION	Redistribution of feature importance: diversity → *EV Driving*, uniqueness → *SmartMeter-1M*	Highlighted key utility drivers dynamically per dataset based on learned importance
Continuous Learning and Feedback	ECHO	Persistent evolution of mapping and weights across runs in response to validation signals	Demonstrated self-correcting adaptability through experience consolidation and feedback integration

Table 5.7 summarizes the key capabilities of AURA across all evaluated datasets. Figure 5.8 illustrates the utility scores computed by AURA for each dataset. The multi-agent reasoning framework within AURA demonstrates quantitative robustness, as evidenced by stable mappings and context-adaptive weights across datasets ranging from 70,000 to 1,000,000 rows. Explainable adaptivity is reflected in interpretable numeric shifts in mappings and weights, which indicate how the agents internalize dataset context. Generalizability is established through consistent performance across transportation, demographic, and time-series domains. Agent cohesiveness results from the

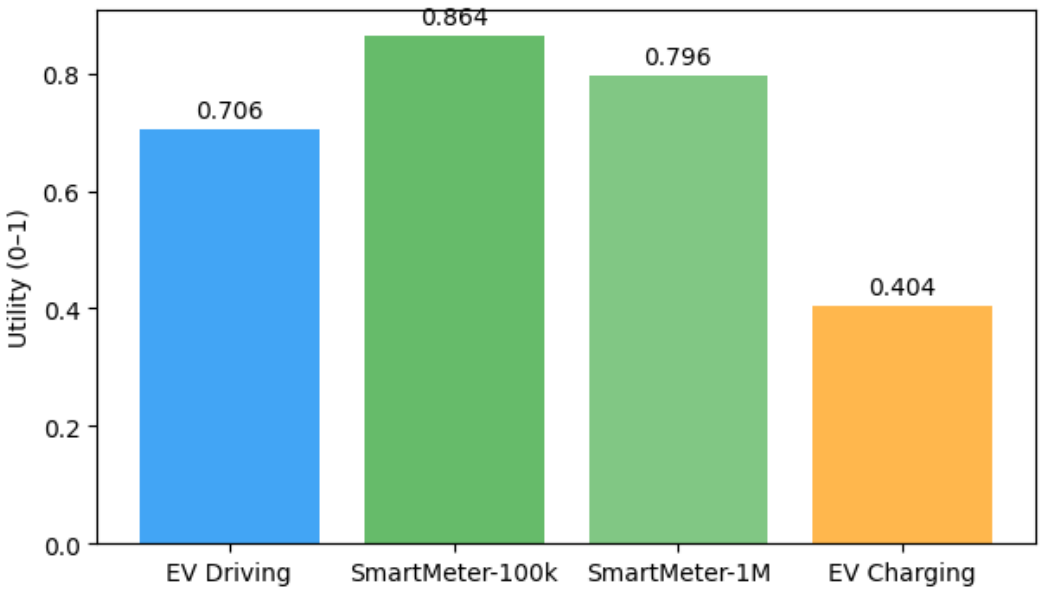


Figure 5.8: AURA: Overall Utility Score Calcualted per Dataset.

integration of QUANTA, CORTEX, FISSION, and ECHO, which collectively transform qualitative perceptions into measurable and continuously improving data-utility assessments.

5.3.2 MAESTRO IMPLEMENTATION ACROSS THE DATASETS

This section provides a comprehensive evaluation of MAESTRO, a target-aware, agentic differential privacy approach, as applied to four heterogeneous datasets. *EV Driving*, *Smart Meter-100K*, *Smart Meter-1M*, and *EV Charging* are evaluated. Table 5.8, informed by Table 5.3, summarizes the intrinsic data utility and composition of each dataset prior to privacy application, as well as the dominant data characteristics influencing utility. Diversity and uniqueness metrics indicate the semantic variability that affects embedding density and sensitivity estimation in subsequent stages.

Each dataset was processed using user-declared privacy-sensitive columns, with $\alpha = 0.4$ set as the maximum augmentation threshold. The evaluation demonstrates how the system autonomously adjusts privacy parameters (ε, δ) and augmentation ratios (α) to meet user-defined data utility targets

Table 5.8: Baseline data utility characteristics of the four datasets.

Dataset	Completeness	Uniqueness	Diversity	Format	Base Utility	Key Driver
EV Driving	1.0	0.04	0.80	1.0	0.706	Diversity
SmartMeter-100K	1.0	1.0	0.83	1.0	0.864	Uniqueness & Diversity
SmartMeter-1M	1.0	1.0	0.80	1.0	0.796	Uniqueness
EV Charging	0.04	0.04	0.76	1.0	0.404	Diversity

while adhering to privacy constraints. Figures 5.9, 5.10, 5.11, and 5.12 present the privacy–utility mapping for each dataset, where MAESTRO iteratively evaluates Laplace and Gaussian mechanisms across an ϵ-grid. The Laplace frontier exhibits greater stability and saturation, especially for the *EV Driving* and *Smart Meter-100K* datasets. In contrast, the Gaussian frontiers show steeper utility decay, supporting MAESTRO's preference for Laplace-based calibration. The ϵ points indicate where each dataset achieves optimal semantic preservation under differential privacy constraints.

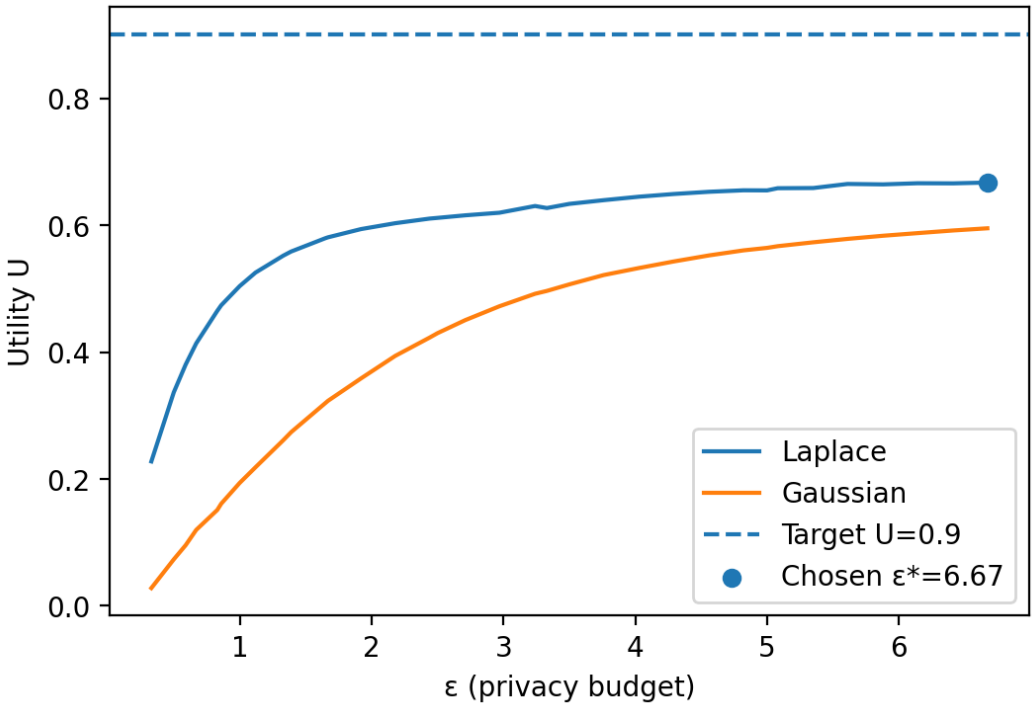


Figure 5.9: Privacy-Utility Mapping for *EV Driving* Dataset

Figure 5.13 quantitatively compares the maximum cosine retention achieved by Laplace and

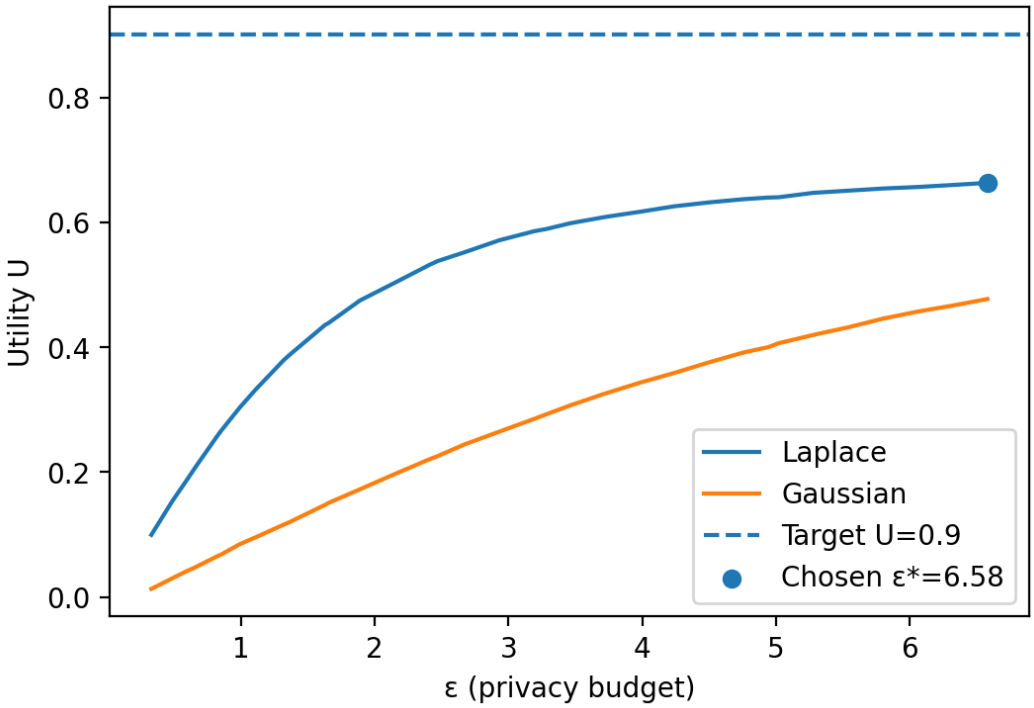


Figure 5.10: Privacy-Utility Mapping for *Smart Meter-100K* Dataset

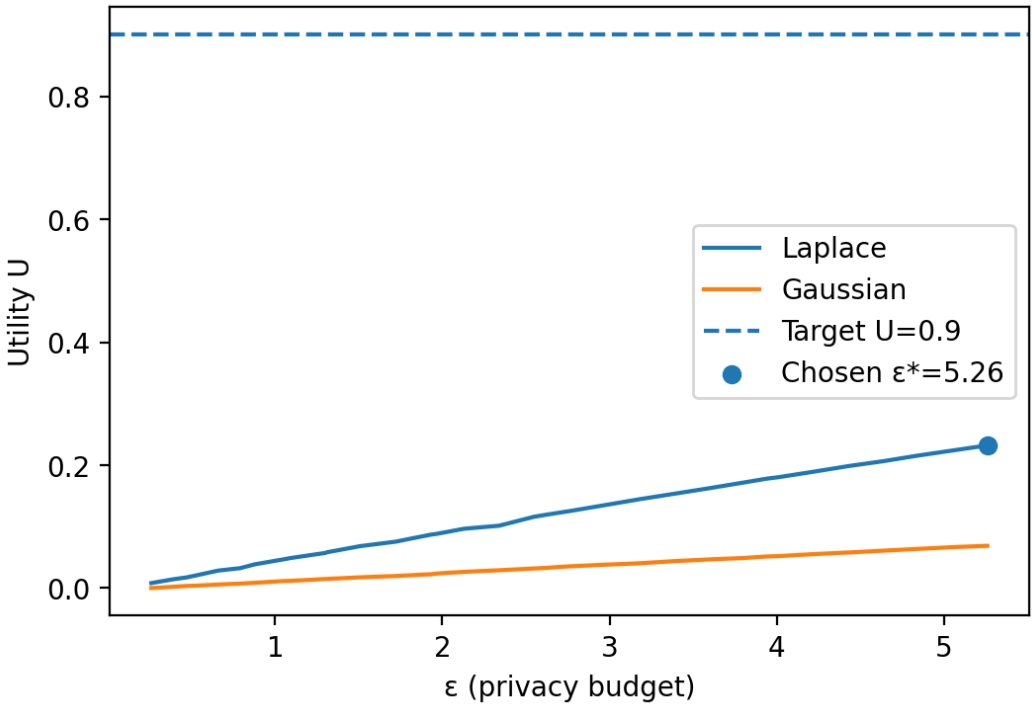


Figure 5.11: Privacy-Utility Mapping for *Smart Meter-1M* Dataset

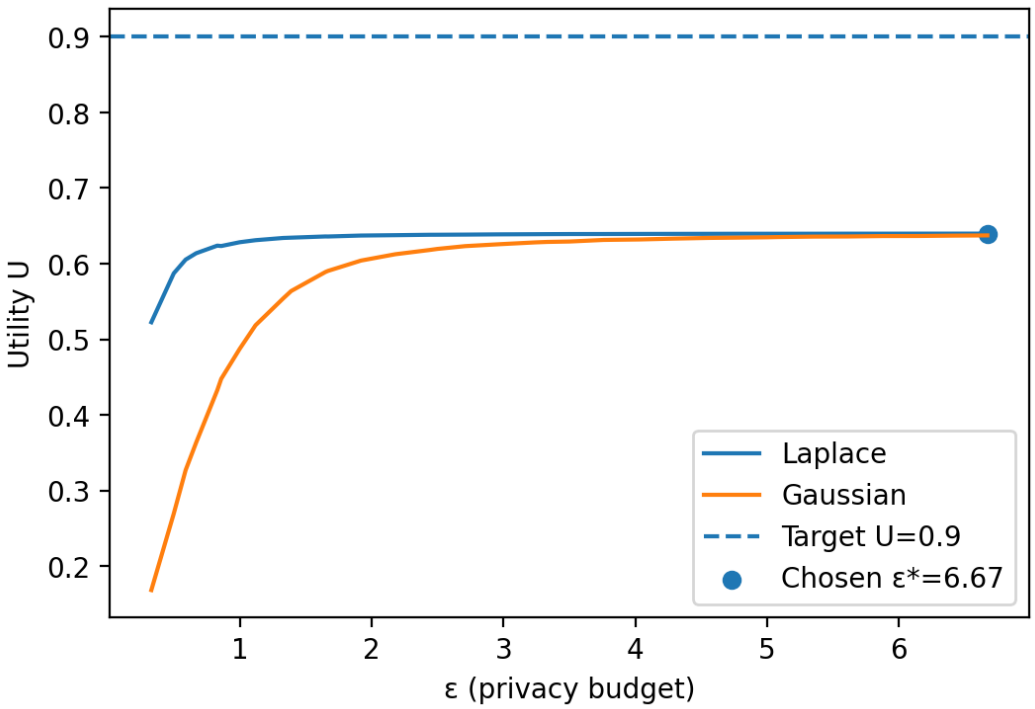


Figure 5.12: Privacy-Utility Mapping for *EV Charging* Dataset

Gaussian mechanisms across all datasets. The Laplace mechanism consistently maintains higher semantic similarity ($u\_cos > 0.9$ for *EV Driving* and *Smart Meter-100K*) even at moderate ϵ values, whereas Gaussian perturbations result in greater semantic degradation. These findings are consistent with theoretical expectations that Laplace noise causes less embedding-space distortion for sparse or categorical text distributions.

The subsequent sections detail how MAESTRO and its sub-agents (SENTINEL, CHIMERA, AEGIS, and ORION) contribute to achieving an adaptive privacy-utility balance.

5.3.2.1 *EV Driving* Dataset

In this dataset, MAESTRO selected the privacy-sensitive columns *age*, *gender*, and *race* based on dataset characteristics. SENTINEL identified low semantic density ($\rho = 0.16-0.50$) and moderate sensitivity ($S \approx 0.63-0.67$) due to the categorical nature of the demographics. AEGIS selected the Laplace mechanism with $\varepsilon^\star = 6.67$, as the application of Gaussian noise resulted in excessive

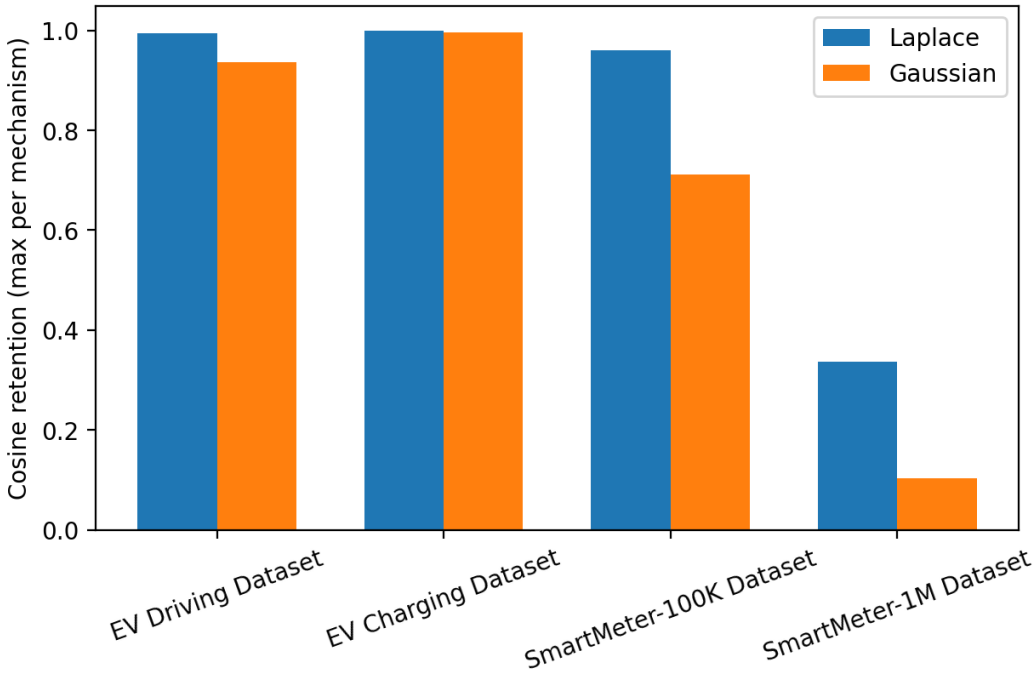


Figure 5.13: Mechanism Comparison (max cosine per mechanism) across the datasets

semantic distortion. CHIMERA's augmentation ($\alpha = 0.4$) stabilized cosine retention ($u_{\cos} \approx 0.99$), while ORION detected moderate cluster deformation ($\Delta\text{sil} \in [-0.08, -0.77]$). MAESTRO ultimately recorded the run as unmet for the strict utility target, but noted strong semantic preservation.

The contributions of the MAESTRO sub-agents are as follows: SENTINEL effectively quantified categorical sensitivity, enabling informed calibration of the Laplace mechanism. CHIMERA augmented context through synonymic demographic expansion, which prevented underfitting of rare categorical pairs. AEGIS balanced privacy and noise, appropriately avoiding the Gaussian mechanism given the low-density categorical embeddings. ORION provided insight into categorical stability using silhouette delta, a critical metric for demographic clustering tasks. MAESTRO coordinated all sub-agents and maintained an ε-preference of 6.67 as an initial calibration for future demographic datasets.

5.3.2.2 *Smart Meter-100K* Dataset

MAESTRO identified *LCLid* as the privacy-sensitive column based on the dataset's characteristics. The dataset exhibited high structural consistency and uniqueness, resulting in a strong baseline utility (0.864). SENTINEL determined minimal sensitivity ($S = 0.11$) and a low semantic density ($\rho = 0.0065$). AEGIS implemented the Laplace mechanism with $\varepsilon^\star = 6.58$, retaining utility at $U = 0.66$ and cosine similarity at $u_{\cos} = 0.96$. The value $\Delta\text{sil} = -0.03$ suggests negligible clustering distortion.

The contributions of the MAESTRO sub-agents are as follows: SENTINEL accurately classified unique IDs as low-sensitivity due to their format. CHIMERA had minimal impact, as text sparsity limited augmentation opportunities. AEGIS effectively achieved the target utility with minimal noise introduction. ORION verified the structural preservation of the data and confirmed numeric resilience. MAESTRO designated this case as a success template for future applications, as high utility was maintained alongside privacy protection.

5.3.2.3 *Smart Meter-1M* Dataset

As with the *Smart Meter-100K* dataset, MAESTRO identified *LCLid* as the privacy-sensitive column. Due to the dataset's large scale and low density ($\rho = 0.00066$), SENTINEL indicated medium sensitivity ($S = 0.26$). AEGIS selected the Laplace mechanism with $\varepsilon^\star = 5.26$; however, overall utility remained limited ($U = 0.23$), and cosine retention (0.34) indicated embedding drift resulting from cumulative noise scaling with N. MAESTRO ultimately classified this as a high-volume, privacy-dominant scenario.

The contributions of the MAESTRO sub-agents are as follows: SENTINEL accurately identified sensitivity inflation caused by scale. As with the *Smart Meter-100K* dataset, CHIMERA had limited impact due to insufficient context diversity at this scale. AEGIS managed differential privacy calibration under massive sample sizes and increased noise amplitude, as required by theoretical

differential privacy bounds. ORION detected rapid utility decay under large-N noise, which is essential for quantifying scale sensitivity. This scenario represents an edge case, demonstrating the scalability limits of fixed-mechanism differential privacy for very large text embeddings.

5.3.2.4 *EV Charging* Dataset

MAESTRO identified *Chargepoint ID* and *Borough* as the privacy-sensitive columns in this case. This dataset exhibited the highest sparsity ($\rho \approx 0.019$) and low completeness (0.04). AEGIS again selected the Laplace mechanism with $\varepsilon^{\star} \approx 4.36{-}4.39$. Although Cosine retention declined moderately ($0.52{-}0.65$), the silhouette change was substantial ($\Delta\text{sil} \approx -0.97$), indicating structural fragility. MAESTRO classified this scenario as failing to meet utility targets despite multiple α trials.

The contributions of the MAESTRO sub-agents are as follows: SENTINEL accurately detected high-risk identification fields, resulting in a higher score ($S = 0.41$). CHIMERA generated station-level contextual paraphrases, but minimal improvement was observed due to limited text coherence. AEGIS reduced ε to enhance privacy, which compromised cluster stability. ORION quantified semantic collapse, which was essential for determining utility lower bounds. MAESTRO ultimately recorded the outcome as privacy-dominant, emphasizing the necessity for feature-specific text augmentation in subsequent analyses.

5.3.2.5 Summary of MAESTRO capabilities and applicability across datasets

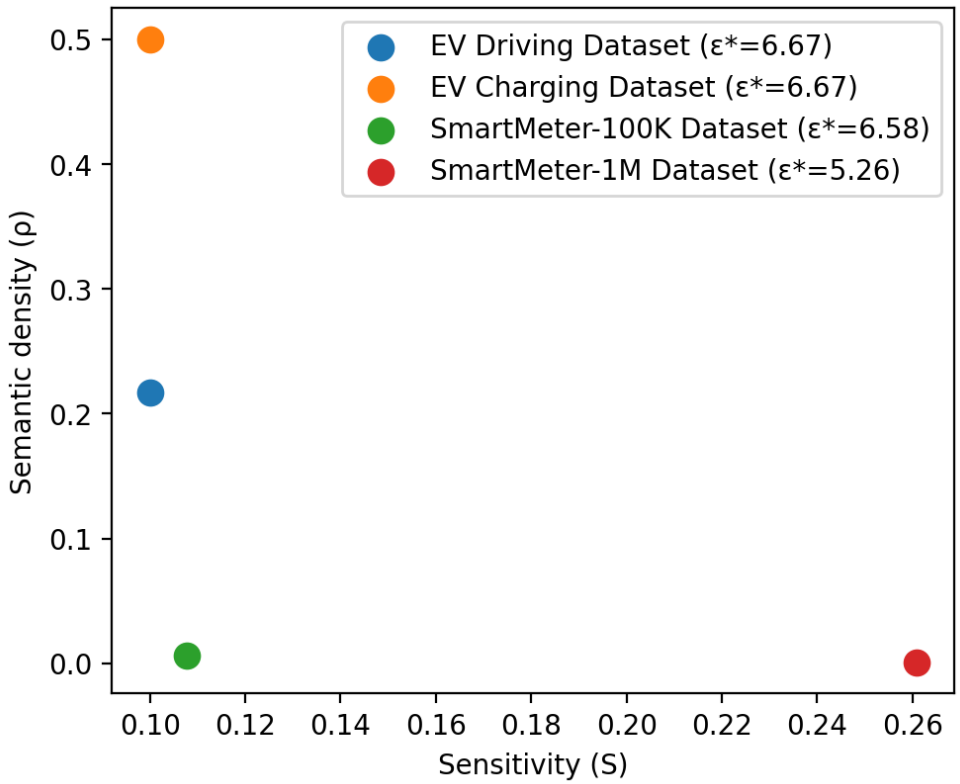


Figure 5.14: Semantic Density-Sensitivity Phase Plot per dataset (marker size $\approx \epsilon^*$)

Analysis across datasets with diverse characteristics indicates that those with higher inherent utility and lower semantic sparsity, such as *Smart Meter-100K*, most closely matched user-specified utility targets (U_target). In contrast, sparse and incomplete datasets, exemplified by the *EV Charging* dataset, revealed that privacy and semantic preservation become adversarial beyond certain thresholds. MAESTRO confirmed that semantic density (ρ) and sensitivity (S) are the primary determinants of feasible privacy levels. Figure 5.14 illustrates the phase relationship between sensitivity (S) and semantic density (ρ) for all datasets, with marker size representing the selected ε. The observed trend shows that datasets with lower semantic density but higher sensitivity, such as the *EV Charging* dataset, required larger ε^* values to achieve even moderate utility targets. Conversely,

datasets with dense, well-structured embeddings, such as the *EV Driving* dataset, attained acceptable utility at smaller ε budgets. This demonstrates the interdependent optimization dynamic between SENTINEL's profiling and AEGIS's parameter calibration.

The results indicate that the augmentation ratio (α) enables only limited recovery of semantic utility and does not fully compensate for high-noise conditions. Dynamic calibration across runs, with parameters stored in memory, stabilizes policy preferences ($\varepsilon_{\text{pref}}$) and accelerates convergence. Figure 5.15 displays a normalized heatmap comparing five metrics: aggregate utility (U), cosine retention (u_cos), silhouette deviation ($|\bigtriangleup sil|$), semantic density (ρ), and sensitivity (S) across the four datasets. The gradient patterns demonstrate that the *EV Driving* and *Smart Meter-100K* datasets maintain high semantic retention despite moderate sensitivity, whereas the *EV Charging* and *Smart Meter-1M* datasets exhibit weaker coherence under differential perturbation. This visualization highlights MAESTRO's cross-dataset consistency and the inverse relationship between sparsity and attainable semantic utility. Table 5.9 consolidates contributions across agents, aligning them with achieved behaviors.

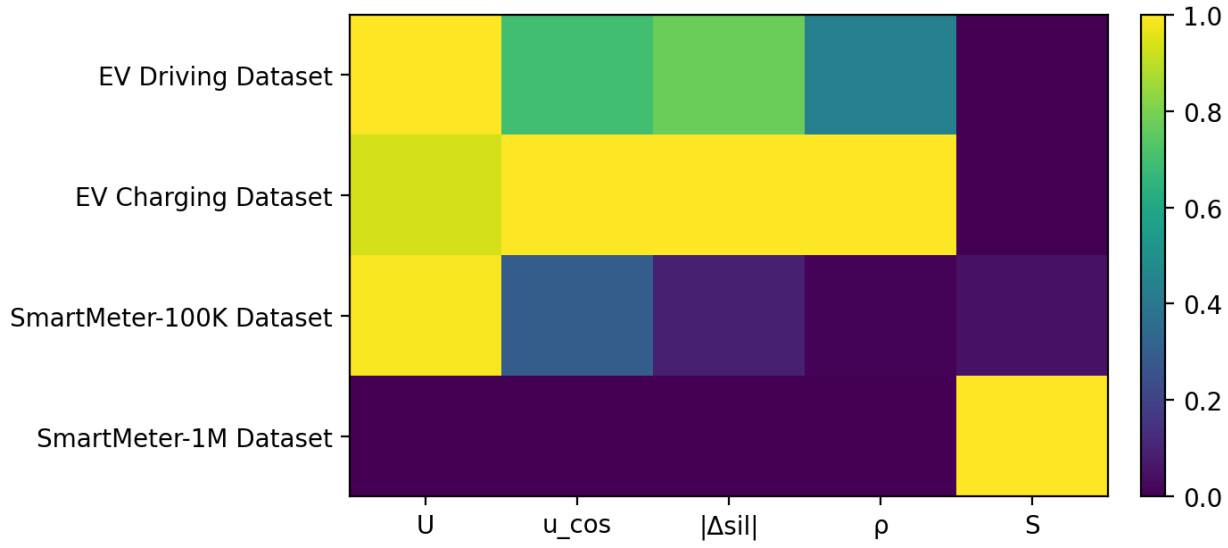


Figure 5.15: Semantic Retention Heatmap (normalized) per dataset

Table 5.9: Summary of MAESTRO characteristics across the four datasets.

Agent	Contribution Evidenced in Experiments
SENTINEL	Accurately mapped user-sensitive columns to embedding-space sensitivity; its PII- and rarity-based scoring provided essential calibration for DP scaling.
CHIMERA	Improved contextual variety for small and categorical datasets, partially offsetting information loss during DP noise injection.
AEGIS	Autonomously selected Laplace mechanism and optimized $\varepsilon^{\star}$ values per dataset; demonstrated adaptive grid refinement.
ORION	Quantified semantic distortion via cosine and silhouette metrics; validated the target-utility satisfaction criterion.
MAESTRO	Orchestrated iterative α-loops, integrated user feedback, and persisted learned parameters for continuous improvement.

5.3.3 Comparative Evaluation: Proposed Framework vs. Non-Private Baseline

To contextualize the privacy-utility trade-offs introduced by the proposed framework, its final outputs are compared against a non-private baseline in which no DP mechanisms are applied. The baseline achieves maximal semantic fidelity, as embeddings remain unperturbed, but offers no formal privacy guarantees. In contrast, the framework enforces explicit (ε, δ)-DP while adaptively optimizing for a user-defined utility target. The following section presents a systematic comparison across four datasets, illustrating how the proposed framework balances semantic utility degradation against quantifiable privacy gains.

Aggregate Utility Comparison Figure 5.16 presents the aggregate utility score U for both the non-private baseline and the proposed framework. The baseline achieves maximal utility, as no perturbation is applied to the embeddings. The framework demonstrates a controlled reduction in utility when calibrated DP noise is introduced. Notably, the achieved utility remains close to the user-specified target threshold, indicating that the framework enforces privacy while preserving a quantifiable level of semantic usefulness. The observed variation across datasets reflects framework's

dataset-aware calibration, which adapts privacy parameters based on sensitivity and semantic density rather than applying a uniform noise budget.

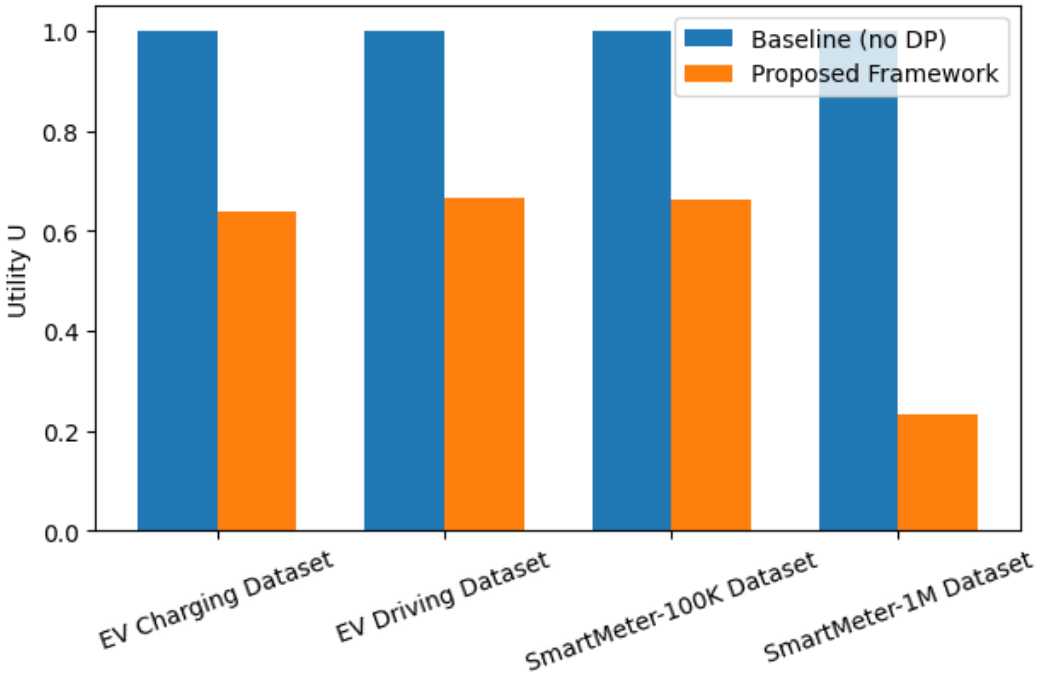


Figure 5.16: Utility (U): Baseline vs MAESTRO

Cosine Retention of Embeddings Figure 5.17 reports cosine retention ($u_{\cos}$), which measures the directional similarity between original and privatized embeddings. The baseline yields perfect cosine retention by construction, whereas the framework exhibits moderate reductions corresponding to the injected noise. Nevertheless, cosine retention remains high across all datasets, indicating that the framework preserves the dominant semantic directions in the embedding space. These findings confirm that the privacy mechanisms selected by AEGIS, guided by SENTINEL's sensitivity estimates, limit semantic distortion rather than arbitrarily perturbing the representation.

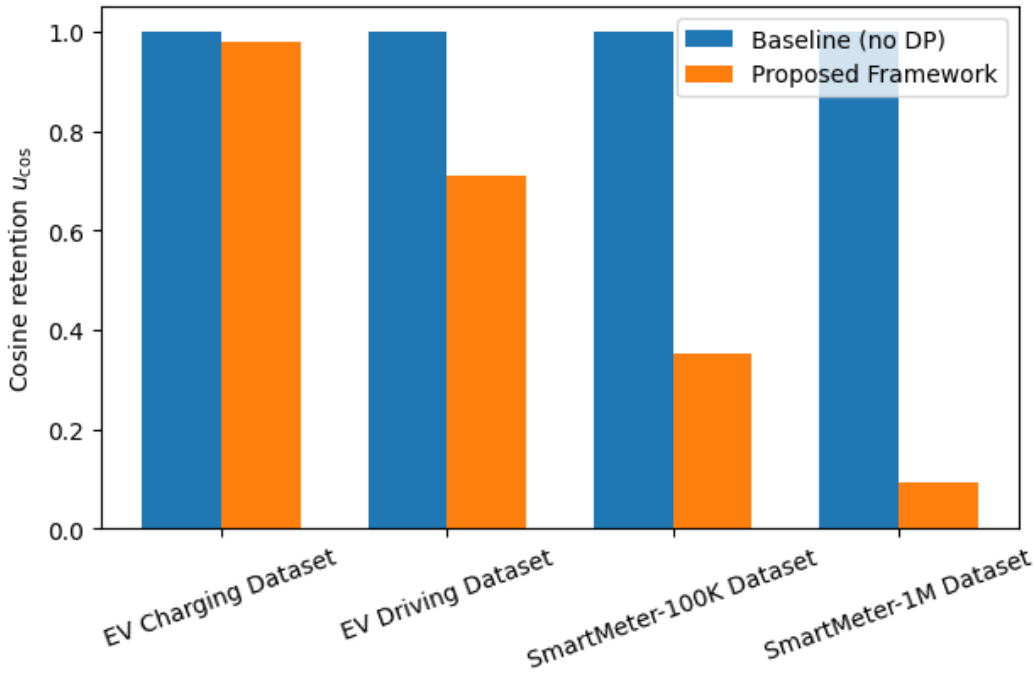


Figure 5.17: Cosine retention: Baseline vs MAESTRO

Structural Change in Embedding Space Figure 5.18 illustrates the silhouette shift Δsil, which captures changes in clustering structure after privacy noise is applied. The non-private baseline shows no structural change, whereas the framework introduces small positive or negative shifts, depending on the dataset. The relatively small magnitude of these shifts indicates that cluster-level semantic organization is largely preserved, even under differential privacy constraints. This result underscores ORION's role in validating that the framework retains both pairwise similarity and higher-level semantic structure.

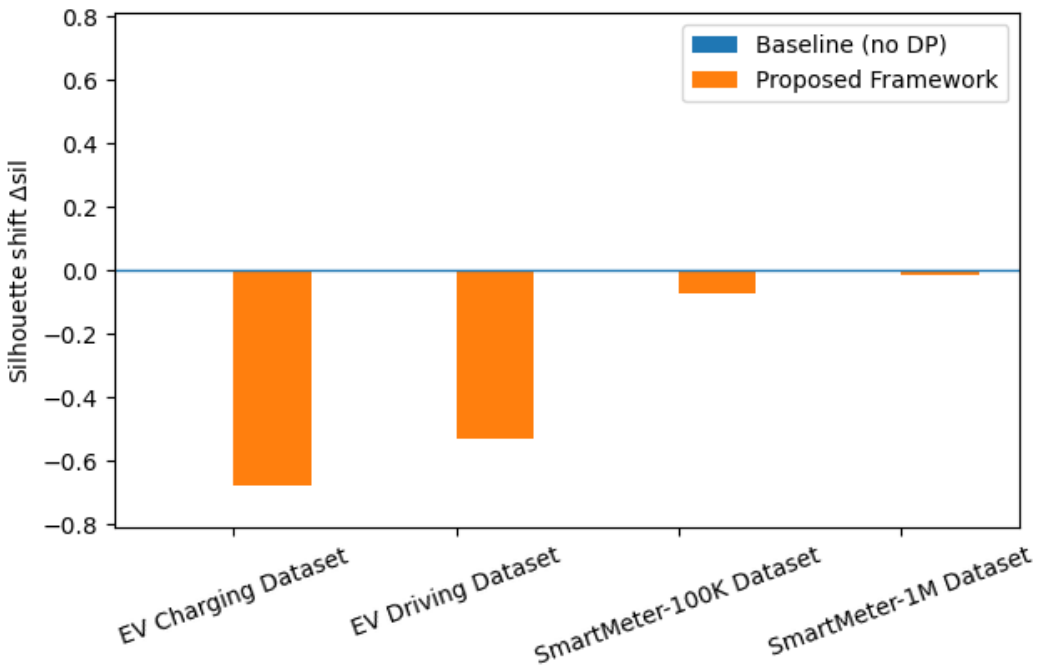


Figure 5.18: Structural change (Δ sil): Baseline vs MAESTRO

Privacy Parameters Selected by the Proposed Framework Figure 5.19 presents the privacy budgets ε^* selected by the framework for each dataset, along with a monotonic privacy-strength proxy defined as $1/(1+\varepsilon^*)$, shown in Figure 5.20. The non-private baseline corresponds to an unbounded privacy budget and therefore provides no formal privacy protection, represented by $\varepsilon \to \infty$. In contrast, the framework outputs finite, dataset-specific ε^* values, reflecting explicit and reportable privacy guarantees. The variation in ε^* across datasets highlights the framework's context-aware optimization: datasets with higher estimated sensitivity or lower semantic density require larger privacy budgets to meet utility targets, whereas less sensitive datasets allow for stronger privacy without excessive utility loss.

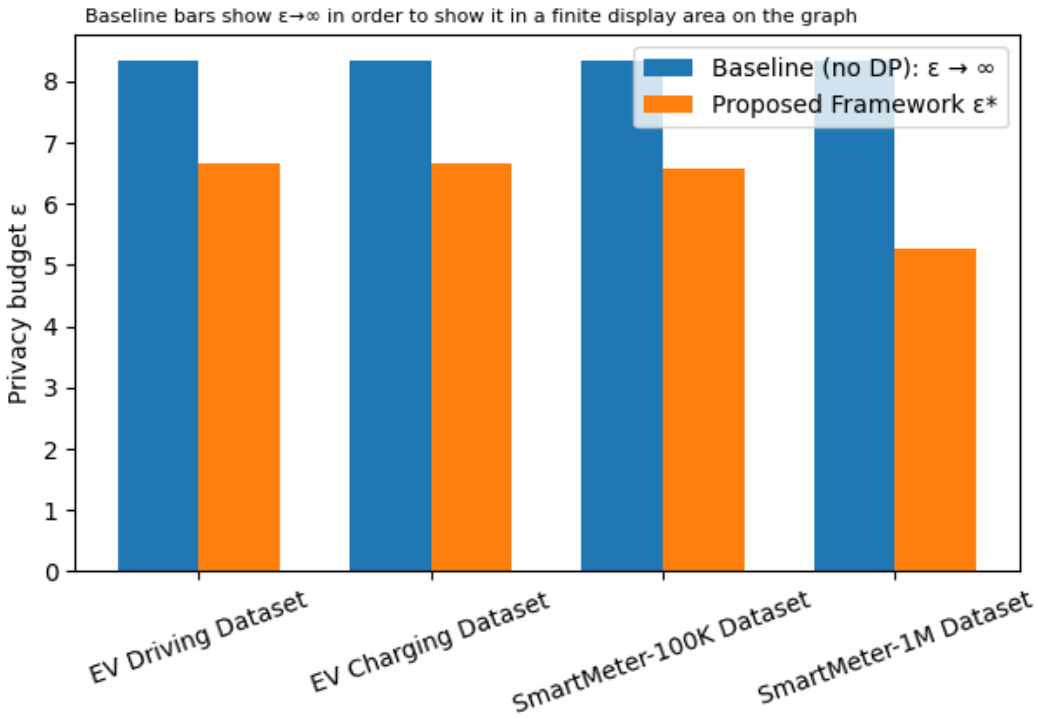


Figure 5.19: MAESTRO chosen privacy budget (lower ϵ = stronger privacy)

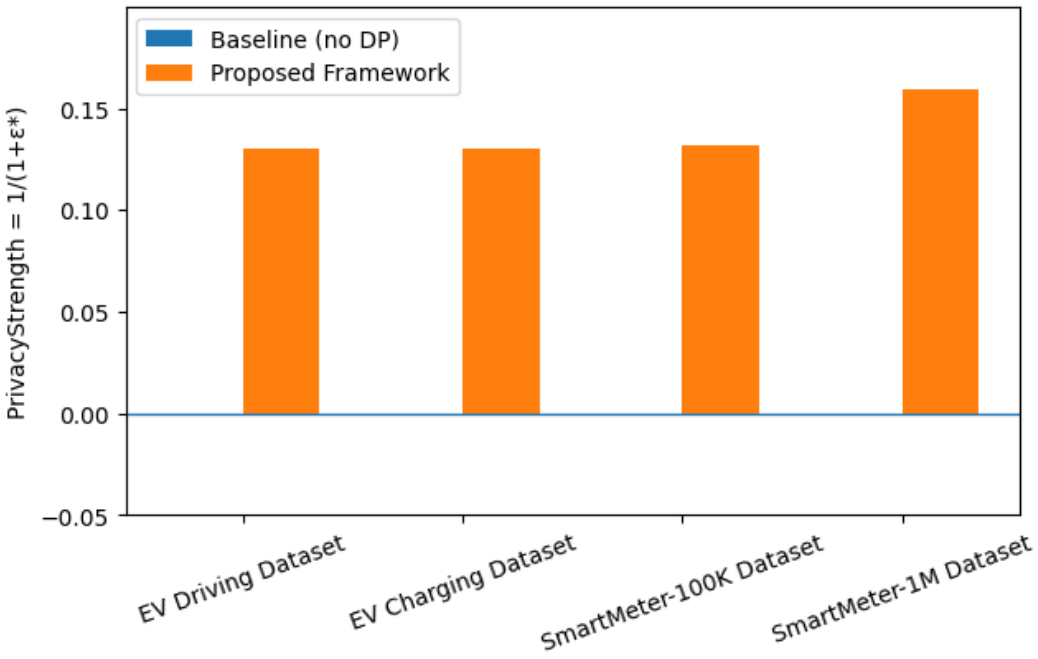


Figure 5.20: Privacy Strength proxy (higher = stronger privacy)

Joint Privacy–Utility Trade-off Figure 5.21 provides a joint view of semantic utility and privacy strength for the framework across all datasets. Each point represents the final operating configuration selected by the framework, with the horizontal axis indicating privacy strength and the vertical axis indicating aggregate utility. This visualization demonstrates that different datasets naturally occupy distinct regions of the privacy–utility space, shaped by their intrinsic sensitivity and semantic characteristics. Rather than enforcing a single global trade-off, the framework identifies a dataset-specific equilibrium that satisfies differential privacy while maintaining quantifiable semantic utility.

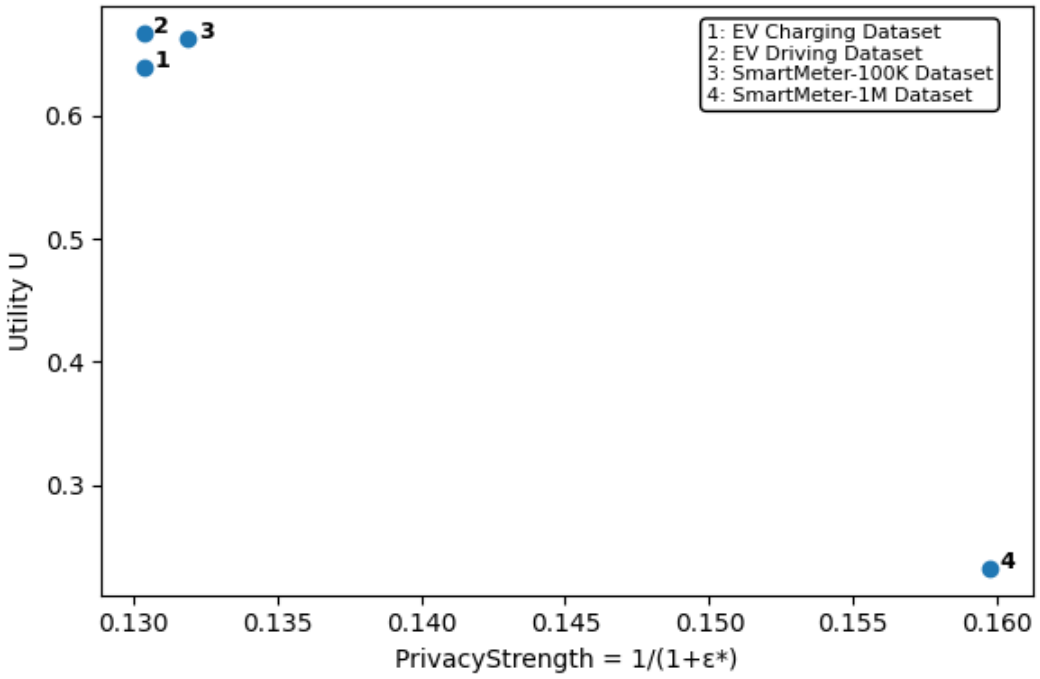


Figure 5.21: Framework trade-off: Utility vs Privacy Strength

Collectively, these results demonstrate that the proposed framework achieves a principled improvement over a non-private baseline by introducing explicit DP guarantees while retaining measurable semantic utility. The comparison indicates that privacy gains are achieved not through indiscriminate noise injection, but through adaptive, dataset-aware calibration informed by profiling, synthesis, and utility feedback. This empirical evidence supports the framework as a systematic

and robust foundational approach for navigating privacy–utility trade-offs in textual embedding pipelines.

5.3.4 Ethical Connotations and Risks of Generative AI in Privacy-Saving Data Exchange

The integration of generative artificial intelligence with privacy-preserving data exchange mechanisms presents significant opportunities alongside complex ethical risks. Although the proposed framework is designed to promote responsible data sharing through adaptive utility reasoning and formal DP guarantees, the agentic and generative components require thorough ethical scrutiny.

A central ethical concern involves the use of generative augmentation mechanisms, such as those implemented by CHIMERA within the MAESTRO workflow. While DP bounds formally restrict individual-level disclosure, generative models may retain higher-order semantic structures that reveal sensitive population-level attributes. This introduces the risk of *semantic over-retention* or more formally *information leakage in latent space*, where group characteristics or latent correlations remain identifiable despite noise injection [140, 242]. The framework addresses this risk by integrating generative augmentation with embedding-aware evaluation (via ORION) and adaptive privacy calibration (via AEGIS), thereby ensuring continuous monitoring of semantic fidelity.

A further ethical consideration involves the automation of privacy-utility trade-offs. By dynamically adjusting privacy budgets (ϵ, δ, and α), MAESTRO alleviates the cognitive demands on end users but centralizes decision-making within algorithmic agents. This centralization introduces the risk of *over-optimization*, where utility objectives may inadvertently drive privacy parameters toward regulatory or ethical limits. The framework mitigates this risk by prioritizing explicit user-defined utility targets, enforcing bounded privacy budgets, and providing transparent reporting of learned parameters to maintain human oversight and accountability.

This study explicitly evaluates the framework's behavior both *with* and *without* DP application. By comparing baseline utility preservation before noise injection to post-privacy performance, the analysis offers an ethically grounded assessment of the trade-offs introduced by privacy mechanisms. This comparative approach contextualizes privacy gains by relating them directly to measurable impacts on utility and semantic fidelity.

5.4 Limitations, Scalability, and Practical Adaptability

5.4.1 General Limitations of the Framework

Despite its modularity and interpretability, the proposed multi-agent framework exhibits inherent limitations. First, the accuracy of qualitative-to-quantitative translation depends critically on the consistency and clarity of user-provided qualitative inputs. While QUANTA learns adaptive mappings over time, ambiguous or contradictory feedback may slow convergence or introduce oscillatory behavior in learned thresholds.

Second, feature importance estimation within FISSION and semantic utility measurement within ORION rely on heuristic and statistical techniques that may not fully capture complex causal relationships among data attributes. In datasets where utility emerges from subtle interactions across features, these agents may underestimate the importance of latent or composite factors. This limitation reflects a broader trade-off between interpretability and expressive power: the framework prioritizes transparent reasoning over opaque, end-to-end optimization.

Although the framework demonstrates the individual performance of each agent through controlled experiments, the evaluation space is necessarily limited. Existing research does not offer comparable end-to-end agentic implementations that address qualitative utility reasoning, contextual adaptation, feature-driven inference, and dynamic privacy calibration simultaneously. Consequently, direct comparisons with existing systems are not feasible. This study therefore adopts a principled

evaluation strategy, comparing baseline performance before and after the application of differential privacy. This approach ensures internal validity while recognizing the limitations present in the broader literature.

5.4.2 Scalability with Extremely Large Datasets

Scalability is a central challenge, especially when the framework is applied to datasets containing millions of records or features with extremely high cardinality. Empirical results indicate that AURA adapts to increasing dataset size by emphasizing diversity and uniqueness, demonstrating contextual sensitivity. However, computational costs increase significantly within MAESTRO, particularly in embedding generation (SENTINEL), clustering-based utility assessment (ORION), and iterative privacy calibration.

Moreover, as dataset size increases, sensitivity estimates tend to rise while semantic density decreases, forcing AEGIS to inject stronger noise to preserve privacy. This phenomenon is explicitly quantified through comparisons between non-private baselines and privacy-preserving executions, revealing the precise points at which privacy constraints begin to dominate utility retention. While this comparison does not replace external benchmarking, it provides a meaningful internal reference that grounds performance claims in measurable trade-offs.

5.4.3 Sparsity and Heterogeneity of Real-World Datasets

Real-world datasets are often sparse, heterogeneous, and partially observed. High levels of missing data can distort context-aware weighting (CORTEX), while heterogeneous column semantics complicate feature importance discovery and sensitivity estimation. In text-heavy datasets, sparsity reduces embedding stability, increasing the volatility of similarity-based utility metrics.

The framework addresses these challenges in part by decoupling intrinsic data characteristics from semantic utility assessment and enabling adaptive reweighting based on contextual signals.

However, sparsity remains a limiting factor in both baseline and privacy-preserving scenarios, underscoring the need to report performance before and after applying differential privacy.

5.4.4 Practical Adjustments and Deployment Strategies

To address the above limitations in practice, the framework supports several pragmatic adjustments. Agents can be selectively activated based on dataset diagnostics, reducing unnecessary computation. Learning rates within ECHO and MAESTRO can be decayed adaptively to prevent instability, while policy snapshots enable rollback and auditability.

From an evaluation perspective, future deployments will benefit from a broader set of quantitative indicators to assess performance. While this study emphasizes utility scores, semantic similarity, and clustering stability, future work will include additional benchmarks, task-specific metrics, and cross-dataset evaluation protocols to better contextualize the effects of privacy mechanisms on performance.

5.5 Summary

This chapter introduced a hierarchical agentic framework for adaptive, privacy-preserving data exchange that integrates human-centered utility reasoning with automated differential privacy actuation. The framework comprises two primary agents: AURA, which translates qualitative user intent and dataset context into interpretable, continuously updated utility objectives, and MAESTRO, which converts these objectives into calibrated privacy mechanisms. Specialized sub-agents support context detection, feature-driven importance inference, semantic and structural utility measurement, mechanism selection, and outcome-driven learning. Implementation workflows detail how the agents interact in a closed loop to sense, reason, act, and learn, enabling dynamic reweighting and policy tuning as data characteristics and user priorities change. Simulation results across multiple datasets demonstrate that this approach preserves downstream utility more effectively than static or

naive privacy baselines while maintaining differential privacy. The chapter concludes by discussing ethical risks, such as governance, misuse, and transparency concerns, as well as practical limitations, including reliance on consistent user feedback, heuristic features, semantic estimators, and scaling costs. Pragmatic deployment strategies for stabilizing learning and enhancing real-world robustness are also outlined.

6 | Conclusion

6.1 Thesis Summary

This thesis investigates responsible and reliable scaling of electrified and autonomous mobility within ecosystems characterized by interconnected energy infrastructure, AI-driven risks, and privacy-sensitive data sharing. Rather than addressing EV adoption, AV risk governance, and privacy-preserving data exchange as separate issues, the thesis presents an integrated narrative that unifies these domains. The large-scale deployment of next-generation mobility depends on three core elements: (i) a robust understanding of adoption trajectories and value streams to inform planning and investment; (ii) lifecycle-based RAI mechanisms to manage safety, security, fairness, and legal risks in autonomy; and (iii) adaptive privacy-utility architectures that facilitate data sharing while minimizing disclosure risks, even in the presence of advanced generative and agentic AI.

These objectives were addressed through three primary research contributions. The first contribution offers a comprehensive analysis of EV proliferation factors, introduces a flexible mathematical model to examine the evolution of EV market share across various jurisdictions and regulatory regimes, and establishes a framework that maps these drivers to both monetary and non-monetary value streams. The second contribution develops a holistic RAI framework for AVs, providing a structured classification of AI risks, including but not limited to safety, security, ethical, and legal domains, and presents detailed bias and fairness identification and mitigation techniques, validated through simulations on publicly available AV datasets. The third contribution introduces an

agentic AI-based framework for textual data that jointly optimizes privacy and utility by translating qualitative user utility expectations into quantitative thresholds, calibrating DP perturbations in embedding space, leveraging responsible generative AI for synthesis and augmentation, and adapting strategies based on dataset profiling and continuous feedback.

Collectively, these research contributions offer both domain-specific advancements and a systems-level perspective on the prerequisites for scalable and trustworthy mobility transformation.

6.1.1 Summary of Research Contribution # 1: EV Proliferation and Value Streams

The first research contribution framed EV adoption as a multi-factor dynamic process influenced by regulatory regimes, technology trajectories, economic considerations, and exogenous influences. It identified and classified EV proliferation factors and proposed a dynamic model to evaluate EV adoption and resulting market share applicable across geographies and jurisdictions. Beyond forecasting adoption, it provided a novel framework to analyze how proliferation drivers interact with emerging value streams. This mapping is strategically important because it connects the adoption problem (how quickly EVs penetrate) to the operational and commercial opportunity space (how EVs can be used to improve grid reliability, deliver environmental benefits, and enable new market roles). The research further emphasized that policy decisions can meaningfully shift adoption trajectories across jurisdictions, even under similar high-level sustainability intentions, as illustrated by policy-relevant analyses and references in the underlying work. Covered in Chapter 3, the contribution is summarized as follows:

- A structured identification of EV proliferation factors and a systematic survey of monetary and non-monetary value streams in the EV domain.
- A dynamic, jurisdiction-agnostic mathematical model for EV adoption that can incorporate regulatory regimes and scenario conditions (including exogenous disruptions) to support broader

applicability beyond a single geography.

- Scenario-driven analysis demonstrating how changes in technology timelines, behavioural shifts, and coordinated factor interventions can alter adoption trajectories.
- A novel framework mapping proliferation factors to value streams, enabling grid operators and market stakeholders to reason about planning timelines and opportunity creation as adoption accelerates.
- A reusable approach for examining policy impacts across jurisdictions, highlighting that policy differences can meaningfully shift EV market share outcomes.

Overall, the first research contribution advances a decision-support perspective, emphasizing that EV proliferation should be modeled not only to predict demand but also to establish timelines for infrastructure readiness and to facilitate coordinated investment decisions that transform adoption challenges into value creation.

6.1.2 Summary of Research Contribution # 2: RAI for AV Lifecycle with a Focus on Bias and Fairness

The second research contribution responded to the reality that AVs are AI-enabled socio-technical systems operating in safety-critical environments. It provided a structured analysis of AI risks in AVs spanning safety hazards, security vulnerabilities, ethical dilemmas, and legal complexities. Building on that risk analysis, the research introduced a holistic RAI framework spanning the AI lifecycle, emphasizing that risks must be identified and mitigated at each stage of AI intervention and that they can compound over time during real-world operation.

Within this framework, the research provides an in-depth examination of bias and fairness risks. It explains how biases can originate during pre-design and data collection, manifest through

modeling and algorithmic choices during design and development, and persist or re-emerge post-deployment due to drift, domain shifts, and evolving operational contexts. Beyond conceptual risk identification, the work proposes practical bias identification and mitigation techniques, validated through simulations using publicly available AV datasets to evaluate strategies such as synthetic data generation and algorithmic fairness analysis. This integration of lifecycle framing and simulation-based methods offers actionable guidance for technologists and stakeholders implementing RAI in AV development. Presented in Chapter 4, the contribution is summarized as follows:

- A structured list and classification of AI risks relevant to AVs spanning nine AI risk domains.
- A holistic RAI framework covering the end-to-end AI lifecycle for AV system development and deployment, emphasizing risk identification, mitigation, and compounding effects across stages.
- Detailed techniques for detecting and mitigating bias and fairness risks across data collection, algorithm design, and real-time decision-making, positioned within the lifecycle framework.
- Simulation-backed demonstrations of bias mitigation strategies using publicly available AV datasets, including evaluation of techniques such as synthetic data generation and algorithmic fairness analysis.

6.1.3 Summary of Research Contribution # 3: Agentic Privacy-Utility Preservation for Text Data

The third research contribution addresses the growing need to share textual data across institutions, systems, and devices, while balancing privacy protection and data utility. Previous approaches often treat utility modeling, privacy calibration, and user intent as separate, static challenges. To overcome these limitations, this research presented an end-to-end, interpretable, agentic architecture that integrates four interconnected components: (1) translating qualitative utility perceptions into quantitative, context-aware thresholds; (2) optimizing DP perturbations in embedding space to

balance semantic fidelity with privacy constraints; (3) leveraging responsible generative AI for data synthesis and augmentation while minimizing sensitive information leakage; and (4) profiling datasets and dynamically adapting privacy strategies based on dataset characteristics and evolving requirements. The contribution is presented in Chapter 5 is summarized as follows:

- An adaptive qualitative-to-quantitative translation mechanism that transforms user-driven qualitative utility perceptions into quantitative utility targets through continual learning.
- Context-aware optimization and feature-driven utility discovery that enable utility scoring to reflect dataset characteristics and user intent rather than relying on fixed metrics.
- A dynamic privacy-utility balancing mechanism that calibrates differential privacy perturbations in embedding space while monitoring semantic fidelity.
- A responsible generative AI approach to synthesis and augmentation that is integrated with privacy calibration and utility evaluation, enabling utility enhancement while managing leakage risk.
- A multi-agent architecture that supports continuous learning and adaptation based on prior executions, outcomes, and user feedback.

Beyond the individual contributions presented in Chapters 3 to 5, this thesis advances a systems-level framework that treats energy-system readiness, responsible autonomy, and privacy-preserving data utility as interdependent requirements for scalable next-generation mobility. Within this framing, EV adoption modeling is directly linked to operational opportunity by clarifying when EV-driven value streams, including grid-relevant opportunities, become practically actionable as adoption grows. The framework also ties operational AI risk governance to data governance by recognizing that fairness and bias in AV systems depend not only on algorithms but also on data availability, representativeness, and governance choices constrained by privacy requirements and the feasibility of cross-stakeholder data sharing. Finally, it positions generative and agentic AI

as dual-purpose capabilities that can enhance system utility, such as through synthetic data for bias mitigation or privacy-preserving augmentation, while simultaneously increasing ethical and governance complexity, thereby necessitating explicit RAI and privacy–utility design.

6.2 Limitations and Challenges

This section consolidates the thesis limitations, encompassing AI ethics (including traditional, generative, and agentic AI), scalability, practical adaptability, implementation challenges, and commercialization barriers. Some limitations are inherent to the research scope, while others represent open challenges that must be addressed for large-scale real-world deployment.

6.2.1 Limitations of the Proposed EV Proliferation Modeling and Value Streams

Modeling Assumptions and Generalization Limits: The EV proliferation model is designed to be reusable across jurisdictions, but any such model necessarily depends on assumptions about which factors are included, how they are parameterized, and how relationships among factors are represented. As proliferation factors evolve over time (e.g., technology cost curves, consumer preferences, infrastructure deployment patterns), coefficients and functional relationships may require re-estimation. The model's applicability, therefore, depends on careful calibration and on the availability and quality of input data for each jurisdictional scenario.

Policy Interpretation and Transferability: The first research contribution highlights that policy differences can shift adoption outcomes and that cross-jurisdictional comparisons can reveal meaningful divergence. However, translating policy instruments into model parameters can be non-trivial. Policies often interact (e.g., incentives plus infrastructure mandates), and implementation effectiveness may differ from legislative intent. As a result, policy transferability across jurisdictions

is limited by local institutional capacity, market structure, and enforcement.

Operationalization of Value Streams: The factor-to-value-stream mapping framework provides a structured way to reason about interdependencies and planning timelines. Nonetheless, realizing value streams in practice often requires additional market rules, interoperability standards, and business model maturity that extend beyond modeling. Some value streams may depend on high-penetration thresholds, aggregator-participation models, or regulatory approvals that vary widely.

6.2.2 Limitations of the Proposed RAI for AVs

Breadth vs. Depth Across Risk Domains: The RAI framework spans multiple AI risk domains and is designed to be adaptable, but the detailed application in the underlying research focuses primarily on bias and fairness. Other risk domains (e.g., certain security threat categories, broader legal compliance mechanisms) are identified but not operationalized and evaluated within the same depth. This creates an opportunity for future work but also delineates a current limitation: the framework's generality does not automatically translate into complete implementations for all domains without additional domain-specific elaboration.

Simulation Limitations and Dataset Representativeness: Bias mitigation simulations provide valuable practical demonstrations, but simulation outcomes depend on the characteristics of the selected datasets, the chosen fairness metrics, and the evaluated mitigation algorithms. Publicly available datasets may not fully represent the diversity of real-world driving environments, geographic variation, sensor configurations, or demographic interactions. Therefore, translating simulation results into real-world confidence requires additional validation in broader operational contexts.

Operational Monitoring and Post-Deployment Governance: Across the AV lifecycle, RAI also emphasizes continuous monitoring, but implementing robust post-deployment governance requires instrumentation, auditing infrastructure, and organizational processes that extend beyond technical model fixes. In practice, continuous monitoring must handle distribution shift, rare events, changing road environments, software updates, and evolving regulatory expectations. These operational realities introduce complexity and cost that can be challenging for commercialization, especially when safety assurance and fairness assurance must be demonstrated simultaneously.

6.2.3 Limitations of the Introduced Agentic Privacy-Utility Framework

Dependence on User Feedback Quality: The third research contribution explicitly relies on qualitative user inputs to shape utility targets and on continual learning to refine mappings. This introduces sensitivity to feedback quality. Ambiguous, inconsistent, or contradictory user feedback can slow convergence or destabilize learned thresholds, and different stakeholders may disagree on what constitutes adequate utility or acceptable privacy.

Differential Privacy and Semantic Fidelity Trade-offs: While differential privacy provides rigorous privacy protection, applying DP to textual embeddings introduces non-trivial challenges to semantic fidelity. Even embedding-aware perturbation strategies can degrade utility for certain tasks, and the relationship between privacy parameters and downstream semantic performance may be dataset-specific. Therefore, the framework's dynamic calibration is valuable, but it cannot eliminate fundamental trade-offs between privacy strength and utility preservation.

Generative AI Risks and Leakage in Latent Structures: The integration of responsible generative AI for synthesis and augmentation can improve dataset utility, but generative methods may retain higher-order semantic structures that risk revealing sensitive population-level attributes even when individual-level disclosure is constrained. The underlying research explicitly notes ethical

connotations and leakage risks that require ongoing scrutiny and careful evaluation. This limitation is not merely technical; it is governance-relevant because stakeholders may require assurances about both individual privacy and group-level inference risks.

Agentic Automation and Accountability: Agentic architectures can centralize decision-making in automated agents (e.g., selecting privacy budgets or strategies) and thereby reduce users' cognitive load. However, automation introduces accountability concerns: if an agent chooses parameters that are near regulatory or ethical limits, responsibility for that decision must be clearly defined. Interpretability helps, but organizational governance is still required to ensure that automated optimization aligns with policy constraints and human oversight expectations.

6.2.4 AI Ethics Limitations Across the Thesis

This thesis engages with AI ethics not as abstract philosophy but as an engineering and governance constraint that shapes adoption, trust, and commercialization.

Traditional AI Ethics (Predictive/Discriminative Models): For traditional AI models, core ethical risks include bias and discrimination, AI opacity or unexplainability, unsafe failures, and misalignment with stakeholder values. The second research contribution directly addresses bias and discrimination risks throughout the AV lifecycle through risk identification and mitigation techniques. Nonetheless, ethical adequacy remains context-dependent. Fairness definitions vary by jurisdiction and stakeholder group, and ethical acceptability may shift as societal expectations evolve. Therefore, no static fairness metric can be assumed universally sufficient.

Generative AI Ethics: Generative AI introduces risks, including hallucination (the generation of plausible but false content), amplification of sensitive attributes, and the potential to generate harmful or misleading outputs. In the thesis, generative AI is used in two primary contexts: (i) as part of the AV bias mitigation toolkit via synthetic data strategies (within the second research

contribution), and (ii) as part of the privacy-utility framework for textual data synthesis and augmentation (the third research contribution). In both contexts, generative outputs can inadvertently encode or leak sensitive patterns. Ethical deployment, therefore, requires systematic monitoring, robust evaluation protocols, and governance controls that are still maturing in practice.

Agentic AI Ethics: Agentic AI introduces additional ethical complexity because agents can make sequences of decisions over time, adapt based on feedback, and optimize objectives that may not fully capture human values. In the third research contribution, agents coordinate privacy and utility decisions and learn over repeated executions. While interpretability and explicit agent roles mitigate some concerns, agentic behaviour can still create unintended consequences (e.g., over-optimizing utility at the boundary of privacy constraints). This creates a need for human-in-the-loop governance, bounded optimization constraints, and explicit accountability for automated decisions.

6.2.5 Scalability Challenges

Computational and Data Scalability: EV proliferation modeling requires data ingestion and calibration across jurisdictions, potentially at a large scale as datasets become richer and more granular. RAI for AV requires continuous monitoring, auditing, and potentially re-training or adaptation across fleets and changing environments. The agentic privacy-utility framework requires repeated profiling, embedding computations, DP calibration, and generative synthesis for large textual corpora. Each scales differently, but all can become computationally intensive, particularly when real-time or near-real-time decisions are required.

Organizational Scalability: Scaling responsible mobility is not purely technical. It requires organizational capabilities: policy alignment, cross-team governance, audit processes, procurement and interoperability, and stakeholder communication. The RAI framework's lifecycle orientation highlights the need for cross-functional coordination (data, machine learning engineering, safety

engineering, legal compliance, product operations), which can be difficult in commercialization contexts due to cost and time-to-market pressures. Similarly, privacy-preserving data exchange across institutions requires governance agreements and trust frameworks beyond algorithm selection.

6.2.6 Practical Adaptability and Implementation Barriers

Integration with Existing Infrastructure and Standards: For EV proliferation, infrastructure readiness is constrained by long planning cycles, regulatory approvals, and physical deployment timelines. Even if adoption is accurately forecasted, implementing timely infrastructure upgrades remains challenging. For AVs, integrating RAI practices with existing AI development pipelines and safety standards is a significant engineering task and requires tooling support for audits, monitoring, and documentation. For privacy-preserving data sharing, implementing DP and agentic optimization in real-world enterprise environments requires integration with data engineering pipelines, access controls, and compliance reporting.

Measurement and KPI Alignment: Practical deployment requires measurable KPIs. EV value streams must be quantifiable to justify investments. RAI metrics for AVs must be meaningful to both engineers and regulators. Privacy-utility metrics must align with user-perceived utility and with formal privacy definitions. The third research contribution explicitly addresses this by translating qualitative utility into quantitative thresholds, but the broader ecosystem challenge remains: aligning stakeholder incentives and measurement frameworks is difficult and context-specific.

6.2.7 Commercialization Challenges

Cost, Time-to-Market, and Competitive Pressure: EV infrastructure upgrades, AV safety assurance, and privacy-preserving data systems all require investment. Competitive markets can pressure organizations to prioritize speed and features over robust governance. RAI for the end-to-end AV lifecycle and continuous monitoring can be seen as overhead unless regulators, insurers, or

customers demand demonstrable assurances. Similarly, privacy-preserving data systems may be under-adopted if benefits are diffuse while implementation costs are immediate.

Regulatory Uncertainty and Liability: Regulatory expectations for AV safety, fairness, and accountability continue to evolve. Liability allocation in AV incidents, standards for fairness demonstration, and audit requirements can change, affecting commercialization risk. For privacy, regulatory regimes differ across jurisdictions and can impose constraints on data sharing. The thesis frameworks are designed to be adaptable, but real-world deployment still faces uncertainty that can delay investment or require redesign.

6.3 Opportunities for Future Work

The contributions of this thesis generate several avenues for future research, both within individual research objectives and at their intersections. These opportunities are best understood in relation to the previously identified limitations and challenges. For EV proliferation modeling and value-stream analysis, a primary direction is systematic model refinement under evolving conditions. The current model relies on assumptions regarding the selection, parameterization, and representation of proliferation factors. Future research should extend the model as technology cost trajectories, consumer preferences, and infrastructure deployment patterns evolve. This extension involves not only incorporating new factors but also re-estimating coefficients, refining functional relationships as the influence of factors shifts, and developing calibration strategies that remain robust despite variations in jurisdiction-specific data quality and availability. Concurrently, strengthening the policy interpretation layer by developing policy mechanism translation toolchains can facilitate systematic mapping of policy instruments to model parameters. This approach addresses the complexity of translating policy into model parameters, particularly when multiple policies interact or when implementation effectiveness diverges from legislative intent. Such toolchains would also support

reproducible comparisons of policy portfolios across jurisdictions and clarify the constraints on policy transferability imposed by local institutional capacity, market structure, and enforcement practices. Another important direction is uncertainty quantification: expanding scenario analysis to include uncertainty bounds and sensitivity analysis will help stakeholders assess the robustness of planning timelines and adoption projections under parameter uncertainty and data variability, which is especially relevant for jurisdiction-agnostic models. Finally, as the realization of value streams is often limited by market rules, interoperability standards, business model maturity, penetration thresholds, aggregator participation models, and regulatory approvals, future research should pursue richer integration with grid operational models. This includes incorporating detailed grid constraints, operational reliability models, and market participation mechanisms to quantify how value streams scale with adoption and to identify the institutional and technical prerequisites for operationalizing them.

Future research on the RAI framework for AVs is shaped by the current trade-off between breadth and depth. While the RAI framework is intended to address multiple AI risk domains, the present work primarily operationalizes bias and fairness. A key direction is to operationalize additional AI risk domains by implementing, evaluating, and benchmarking mitigation strategies beyond bias and fairness, including specific security threats and broader legal compliance mechanisms. Advancing in this area requires stronger validation under real-world representativeness constraints. Current bias mitigation simulations depend on specific datasets and metrics, and publicly available datasets may not capture the full diversity of driving environments, geographic variation, sensor configurations, or demographic interactions. Therefore, future research should broaden validation to more diverse operational contexts and develop evaluation methods that explicitly address representativeness limitations. Additionally, the framework's lifecycle orientation requires investment in post-deployment auditing and monitoring systems. Developing scalable, standardized monitoring pipelines to assess fairness and other risks under distribution shifts is essential, given operational realities such as rare events, changing road environments, software updates, and evolving regulatory

requirements. These factors also contribute to implementation complexity and cost, intensifying commercialization pressures when safety and fairness must be demonstrated concurrently. To facilitate comparability and industry adoption, future work should develop standardized evaluation suites that offer reproducible benchmarks for AV fairness and broader RAI compliance. Furthermore, governance integration can be advanced by defining organizational templates that embed RAI principles throughout the AV lifecycle and product development process. These templates should include documentation practices, audit trails, and accountability structures aligned with regulatory expectations, thereby addressing the need for comprehensive organizational processes beyond technical model adjustments.

Future research on the agentic privacy-utility framework is shaped by previously identified sources of sensitivity and risk. As the framework depends on qualitative user feedback to define utility targets and employs continual learning to refine mappings, a primary direction is to enhance the robustness of intent capture and feedback handling. This will ensure that ambiguous, inconsistent, or contradictory feedback does not destabilize learned thresholds and that disagreements among stakeholders regarding acceptable privacy and utility levels are explicitly managed. Simultaneously, further investigation into the trade-off between differential privacy and semantic fidelity is warranted. Embedding-aware perturbation may reduce utility for certain tasks, and the relationship between privacy parameters and downstream semantic performance is often dataset-specific. Therefore, future work should broaden evaluation across tasks and domains using additional quantitative indicators and standardized benchmarks to contextualize improvements. Methodological advancements can be achieved by exploring alternative noise mechanisms and privacy strategies, including various differential privacy mechanisms, alternative privacy definitions, and hybrid approaches, to improve adaptability and performance across different dataset types while clarifying the interaction between privacy choices and utility degradation. Given the integration of responsible generative AI for synthesis and augmentation, future research should also systematically evaluate leakage risks in latent structures, providing governance-relevant assurances regarding both individual-level privacy

and sensitive population-level attribute inference. Scaling the architecture for large datasets and enterprise applications necessitates distributed agent execution, including strategies for distributed execution, orchestration, and cost-aware optimization to address the computational demands of repeated profiling, embedding computations, differential privacy calibration, and generative synthesis. Extending the framework beyond text to non-tabular and multimodal data types, such as images, graphs, and multimodal corpora, remains a significant direction. Achieving this will require corresponding adaptations in utility metrics and privacy mechanisms to maintain the framework's intent-aware, context-dependent optimization capabilities.

These objective-specific directions also imply a set of cross-cutting opportunities at the intersection of the three research contributions, particularly because several limitations are inherently systemic. One such opportunity is *privacy-preserving data sharing for mobility and energy planning*. EV adoption forecasting and value-stream realization can benefit from richer data sharing across utilities, OEMs, fleets, and regulators, yet data sharing is constrained by privacy and confidentiality concerns. The privacy-utility architecture from the third research contribution, therefore, suggests a pathway to share sensitive operational text and logs while preserving privacy, thereby improving calibration quality and planning robustness in the EV modeling work. A second intersectional opportunity concerns *synthetic data governance across fairness and privacy*. Synthetic data is relevant to both AV fairness mitigation and privacy-preserving data exchange. Future work can develop unified governance and evaluation protocols for synthetic data that jointly address fairness improvements and leakage risks, explicitly combining lifecycle RAI considerations with privacy-utility metrics rather than treating these as separate evaluation tracks. A third opportunity is *responsible agentic optimization in safety-critical systems*. Because agentic AI introduces ethical complexity, agents can adapt over time, optimize objectives that may not fully capture human values, and potentially over-optimize at the boundary of constraints. Future work can explore bounded and auditable agentic optimization techniques that maintain human oversight and comply with safety and fairness constraints in AV settings, thereby directly addressing accountability concerns raised

by agentic automation. Finally, an integrative direction is the creation of *decision-support systems for policymakers and operators* that combine EV adoption models, RAI lifecycle assessments, and privacy-preserving analytics into unified platforms. Such platforms would help stakeholders evaluate trade-offs among adoption acceleration, safety and fairness assurance, and privacy protection within a consistent framework, and would also operationalize measurement and KPI alignment across domains by connecting technical metrics to planning, governance, and regulatory decision-making needs.

6.4 Final Remarks

The mobility transition constitutes both a technological shift and a governance and systems-integration challenge. The first research contribution demonstrates that EV proliferation can be accelerated through coordinated interventions and that mapping proliferation drivers to value streams enables opportunity-driven planning. The second research contribution establishes that AV deployment requires lifecycle-based responsible AI mechanisms, with bias and fairness risks necessitating actionable detection, mitigation, and evaluation strategies grounded in practice. The third research contribution illustrates that privacy and utility can be addressed as complementary objectives through adaptive, agentic designs that integrate differential privacy, responsible generative AI, and continual learning.

Collectively, these contributions establish a foundation for advancing electrified and autonomous mobility as a trustworthy socio-technical system, capable of scaling while maintaining grid reliability, ensuring responsible AI behavior in safety-critical autonomy, and enabling data sharing that respects privacy and preserves utility. The future work opportunities identified throughout the thesis suggest a path toward deeper integration of energy modeling, responsible autonomy, and privacy-preserving data governance, which is increasingly necessary as mobility systems become more electrified, autonomous, and data-driven.

Bibliography

[1] International-Energy-Agency, *Energy Technology Policy Division of the Directorate of Sustainability, Technology and Outlooks*, 2020. [Online]. Available: https://www.iea.blob.core.windows.net/assets/7f8aed40-89af-4348-be19-c8a67df0b9ea/Energy_Technology_Perspectives_2020_PDF.pdf. [Accessed: 26 January 2026].

[2] Global-EV-Outlook, *Energy Technology Policy Division of the Directorate of Sustainability, Technology and Outlooks*, International Energy Agency, 2019. [Online]. Available: https://www.iea.org/topics/energy-technology-perspectives. [Accessed: 26 January 2026].

[3] B. Crothers, "This Chinese City Has 16,000 Electric Buses And 22,000 Electric Taxis," Forbes, 2021. [Online]. Available: https://www.forbes.com/sites/brookecrothers/2021/02/14/this-chinese-city-has-16000-electric-buses-and-22000-electric-taxis/?sh=464c86f63a92. [Accessed: 26 January 2026].

[4] K. Larsen, "Tesla taxi hits the road in Vancouver | CBC News," *CBCnews*, CBC/Radio Canada, 2020, Sep. [Online]. Available: https://www.cbc.ca/news/canada/british-columbia/tesla-taxi-hits-the-road-in-vancouver-1.5737062. [Accessed: 26 January 2026].

[5] K. Monks, "There's a new entry in India's electric rickshaw race," CNN Business, 2020. [Online]. Available: https://www.cnn.com/2020/03/25/energy/altigreen-india-electric-rickshaw-spc-intl/index.html. [Accessed: 26 January 2026].

[6] T. Ward, "Walmart Teams Up with Cruise to Pilot All-Electric Self-Driving Delivery Powered by 100% Renewable Energy," Walmart, 2020. [Online]. Available: https://corporate.walmart.com/newsroom/2020/11/10/walmart-teams-up-with-cruise-to-pilot-all-electric-self-driving-delivery-powered-by-100-renewable-energy. [Accessed: 26 January 2026].

[7] SEPA, "Preparing for an Electric Vehicle Future: How Utilities Can Succeed," Smart Electric Power Alliance, 2019. [Online]. Available: https://sepapower.org/resource/preparing-for-an-electric-vehicle-future-how-utilities-can-succeed/. [Accessed: 26 January 2026].

[8] E. Schmidt, "EV clustered charging can be problematic for electrical utilities," FLEET CARMA, 2017. [Online]. Available: https://www.fleetcarma.com/ev-clustered-charging-can-problematic-electrical-utilities/. [Accessed: 26 January 2026].

[9] A. Brown, S. Lommele, A. Schayowitz, and and E. Klotz, "Electric Vehicle Charging Infrastructure Trends from the Alternative Fueling Station Locator: First Quarter 2020," National Renewable Energy Laboratory, 2020. [Online]. Available: https://www.nrel.gov/docs/fy20osti/77508.pdf. [Accessed: 26 January 2026].

[10] IEA, *World Energy Investment*, International Energy Agency, 2021. [Online]. Available: https://iea.blob.core.windows.net/assets/5e6b3821-bb8f-4df4-a88b-e891cd8251e3/WorldEnergyInvestment2021.pdf. [Accessed: 26 January 2026].

[11] C. Hanvey, "EV Managed Charging: Lessons from Utility Pilot Programs," Smart Electric Power Alliance, 2019. [Online]. Available: https://sepapower.org/knowledge/ev-managed-charging-lessons-from-utility-pilot-programs/. [Accessed: 26 January 2026].

[12] M. Woodward, Dr. B. Walton, Dr. J. Hamilton, G. Alberts, S. Fullerton-Smith, E. Day, and J. Ringrow, "Electric vehicles: Setting a course for 2030," Deloitte Insights, 2020. [Online]. Available: https://www2.deloitte.com/us/en/insights/focus/future-of-mobility/electric-vehicle-trends-2030.html. [Accessed: 26 January 2026].

[13] M. Nicholas, D. Hall, and N. Lutsey, "QUANTIFYING THE ELECTRIC VEHICLE CHARGING INFRASTRUCTURE GAP ACROSS U.S. MARKETS," ICCTV, The International Council on Clean Transportation, 2019. [Online]. Available: https://theicct.org/publications/charging-gap-US. [Accessed: 26 January 2026]. [Accessed: 26 January 2026].

[14] M. Melaina, B. Bush, J. Eichman, E. Wood, D. Stright, V. Krishnan, D. Keyser, T. Mai, J. Mclaren, "National Economic Value Assessment of Plug-In Electric Vehicles Volume I," 2016, doi: 10.13140/RG.2.2.26728.98563. [Online]. Available: https://doi.org/10.13140/RG.2.2.26728.98563.

[15] Y. Zhang, M. Zhong, N. Geng, Y. Jiang, "Forecasting electric vehicles sales with univariate and multivariate time series models: The case of China," *PLOS ONE*, vol. 12, no. 5, pp. e0176729, May, 2017, doi: 10.1371/journal.pone.0176729. [Online]. Available: https://doi.org/10.1371/journal.pone.0176729.

[16] B. Feng, Q. Ye, B. J. Collins, "A dynamic model of electric vehicle adoption: The role of social commerce in new transportation," *Information & Management*, vol. 56, no. 2, pp. 196–212, Mar, 2019, doi: 10.1016/j.im.2018.05.004. [Online]. Available: https://doi.org/10.1016/j.im.2018.05.004.

[17] V. Bilotkach, M. Mills, "Simple Economics of Electric Vehicle Adoption," *Procedia - Social and Behavioral Sciences*, vol. 54, pp. 979–988, Oct, 2012, doi: 10.1016/j.sbspro.2012.09.813. [Online]. Available: https://doi.org/10.1016/j.sbspro.2012.09.813.

[18] G. H. Broadbent, D. Drozdzewski, G. Metternicht, "Electric vehicle adoption: An analysis of best practice and pitfalls for policy making from experiences of Europe and the US," *Geography Compass*, vol. 12, no. 2, pp. e12358, Dec, 2017, doi: 10.1111/gec3.12358. [Online]. Available: https://doi.org/10.1111/gec3.12358.

[19] P. Krutko, J. C. Moon, and J. A. Finkle, "Analysis of the Electric Vehicle Industry," International Economic Development Council, IEDC, 2013. [Online]. Available: https://www.iedconline.org/clientuploads/Downloads/edrp/IEDC_Electric_Vehicle_Industry.pdf. [Accessed: 26 January 2026].

[20] E. Kim, E. Heo, "Key Drivers behind the Adoption of Electric Vehicle in Korea: An Analysis of the Revealed Preferences," *Sustainability*, vol. 11, no. 23, pp. 6854, Dec, 2019, doi: 10.3390/su11236854. [Online]. Available: https://doi.org/10.3390/su11236854.

[21] I. Malmgren, "Quantifying the Societal Benefits of Electric Vehicles," *World Electric Vehicle Journal*, vol. 8, no. 4, pp. 996–1007, Dec, 2016, doi: 10.3390/wevj8040996. [Online]. Available: https://doi.org/10.3390/wevj8040996.

[22] NRDC, "Scaling Up Electric Vehicle Charging Infrastructure," Natural-Resources Defense Council, 2020. [Online]. Available: https://www.nrdc.org/sites/default/files/charging-infrastructure-best-parctices-202007.pdf. [Accessed: 26 January 2026].

[23] R. Dua, S. Hardman, Y. Bhatt, D. Suneja, "Enablers and disablers to plug-in electric vehicle adoption in India: Insights from a survey of experts," *Energy Reports*, vol. 7, pp. 3171–3188, November, 2021, doi: 10.1016/j.egyr.2021.05.025. [Online]. Available: https://doi.org/10.1016/j.egyr.2021.05.025.

[24] A. Soman, K. Ganesan, H. Kaur, "India's Electric Vehicle Transition: Impact on Auto Industry and Building the EV Ecosystem.," 2019. [Online]. Available: https://www.ceew.in/sites/default/files/CEEW-IndiaElectricVehicleTransitionReportPDF26Nov19.pdf. [Accessed: 26 January 2026].

[25] I. Wagner, "Number of public electric vehicle charging stations and charging outlets in the U.S. as of February 16, 2021," Statista, 2021. [Online]. Available: https://www.statista.com/statistics/416750/number-of-electric-vehicle-charging-stations-outlets-united-states/. [Accessed: 26 January 2026].

[26] I. Wagner, "Number of gasoline station establishments in the United States from 2013 to 2016," Statista, 2020. [Online]. Available: https://www.statista.com/statistics/525107/number-of-gasoline-stations-in-the-united-states/. [Accessed: 26 January 2026].

[27] Market-Watch-Inc, "How Many Gas Stations Are In U.S.? How Many Will There Be In 10 Years?," MarketWatch, 2020. [Online]. Available: https://www.marketwatch.com/story/how-many-gas-stations-are-in-us-how-many-will-there-be-in-10-years-2020-02-16. [Accessed: 26 January 2026].

[28] StoreDot, "Extreme-Fast Charging Technology: Taking EV charging from hours to minutes," 2021. [Online]. Available: https://www.store-dot.com. [Accessed: 26 January 2026].

[29] D. Carrington, "Electric car batteries with five-minute charging times produced," The Guardian, The Guardian, 2021. [Online]. Available: https://www.theguardian.com/environment/2021/jan/19/electric-car-batteries-race-ahead-with-five-minute-charging-times. [Accessed: 26 January 2026].

[30] Chargepoint, "The Electric Revolution Is Here," 2021. [Online]. Available: https://www.chargepoint.com/en-ca/businesses/industries/. [Accessed: 26 January 2026].

[31] Flo, "FLO is a leading North American charging network," 2021. [Online]. Available: https://www.flo.com/en-CA/. [Accessed: 26 January 2026].

[32] Tesla, "Tesla Wall Connector and Superchargers," 2021. [Online]. Available: https://www.tesla.com/en_CA/support/home-charging-installation/wall-connector. [Accessed: 26 January 2026].

[33] Z. Wang, X. Wang, L. Wang, X. Hu, W. Fan, "Research on electric vehicle (EV) driving range prediction method based on PSO-LSSVM," pp. 260-265, 2017, doi: 10.1109/ICPHM.2017.7998338. [Online]. Available: https://doi.org/10.1109/ICPHM.2017.7998338.

[34] P. Valdes-Dapena, "Electric car batteries are catching fire and that could be a big turnoff to buyers," 2020. [Online]. Available: https://www.cnn.com/2020/11/10/success/electric-car-vehicle-battery-fires/index.html. [Accessed: 26 January 2026].

[35] Battery-University, "BU-1003a: Battery Aging in an Electric Vehicle (EV)s," 2021. [Online]. Available: https://batteryuniversity.com/learn/article/bu_1003a_battery_aging_in_an_electric_vehicle_ev. [Accessed: 26 January 2026].

[36] Geotab, "Electric Vehicle Battery Degradation Tool," 2021. [Online]. Available: https://www.geotab.com/fleet-management-solutions/ev-battery-degradation-tool/. [Accessed: 26 January 2026].

[37] Tesla, "Model Y Achieves 5-Star Overall Safety Rating from NHTSA," 2021. [Online]. Available: https://www.tesla.com/en_CA/blog/model-y-achieves-5-star-overall-safety-rating-nhtsa. [Accessed: 26 January 2026].

[38] C. Nguyen, "Why Tesla's Model 3 received top crash-test safety ratings," 2020. [Online]. Available: https://www.businessinsider.com/why-tesla-model-3-received-5-star-crash-test-rating-2019-10. [Accessed: 26 January 2026].

[39] Volkwagen, "Battery Safety QA: Electric Car information," Volkswagen UK, 2021. [Online]. Available: https://www.volkswagen.co.uk/en/electric-and-hybrid/software-and-technology/battery-technology/battery-safety.html. [Accessed: 26 January 2026].

[40] I. Boudway, "Batteries For Electric Cars Speed Toward a Tipping Point," Bloomberg Hyperdrive, Bloomberg Finance L.P., 2020, Dec. [Online]. Available: https://www.bloomberg.com/news/articles/2020-12-16/electric-cars-are-about-to-be-as-cheap-as-gas-powered-models. [Accessed: 26 January 2026].

[41] BloombergNEF, "Battery Pack Prices Fall As Market Ramps Up With Market Average At $156/kWh In 2019," Bloomberg Finance L.P., 2019. [Online]. Available: https://about.bnef.com/blog/battery-pack-prices-fall-as-market-ramps-up-with-market-average-at-156-kwh-in-2019. [Accessed: 26 January 2026].

[42] M. Dent, "FORECAST: EV copper demand to rise 9-fold by 2027," 2017. [Online]. Available: https://www.metalbulletin.com/Article/3726147/FORECAST-EV-copper-demand-to-rise-9-fold-by-2027.html. [Accessed: 26 January 2026].

[43] J. Ziebart, "Putting the Copper Horse Before the EV Cart: Copper Demand in the EV Market," 2018. [Online]. Available: https://investingnews.com/innspired/the-electric-vehicle-market-and-copper-demand/. [Accessed: 26 January 2026].

[44] The-White-House, "FACT SHEET: President Biden Sets 2030 Greenhouse Gas Pollution Reduction Target Aimed at Creating Good-Paying Union Jobs and Securing U.S. Leadership on Clean Energy Technologies," 2021. [Online]. Available: https://www.whitehouse.gov/briefing-room/statements-releases/2021/04/22/fact-sheet-president-biden-sets-2030-greenhouse-gas-pollution-reduction-target-aimed-at-creating-good-paying-union-jobs-and-securing-u-s-leadership-on-clean-energy-technologies/. [Accessed: 26 January 2026].

[45] European-Commission, "CO2 emission performance standards for cars and vans," 2020. [Online]. Available: https://ec.europa.eu/clima/policies/transport/vehicles/regulation_en. [Accessed: 26 January 2026].

[46] Climate-Analytics, "Climate Action Tracker," 2021. [Online]. Available: https://climateactiontracker.org/countries/india/pledges-and-targets/. [Accessed: 26 January 2026].

[47] C2ES, "U.S. State Greenhouse Gas Emissions Targets," Center for Climate and Energy Solutions, 2021. [Online]. Available: https://www.c2es.org/document/greenhouse-gas-emissions-targets/. [Accessed: 26 January 2026].

[48] The-Associated-Press, "London taxes older vehicles in bid to fight air pollution," 2019. [Online]. Available: https://www.ctvnews.ca/autos/london-taxes-older-vehicles-in-bid-to-fight-air-pollution-1.4370483. [Accessed: 26 January 2026].

[49] MTO, "High Occupancy Vehicle (HOV) Lanes," Ministry of Tranportation, Ontario, 2021. [Online]. Available: http://www.mto.gov.on.ca/english/ontario-511/hov-lanes.shtml. [Accessed: 26 January 2026].

[50] S. Ardiyok, A. Canbeyli, and J. Skardziuteo, "Turkey: How Europe Promotes Electric Vehicles?: A Brief Insight On Best Practices," 2020. [Online]. Available: https://www.mondaq.com/turkey/rail-road-cycling/904350/how-europe-promotes-electric-vehicles-a-brief-insight-on-best-practices-. [Accessed: 26 January 2026].

[51] Vinci-Group, “Are self-driving cars about to get their own lane?,” 2019. [Online]. Available: https://leonard.vinci.com/en/are-self-driving-cars-about-to-get-their-own-lane/. [Accessed: 26 January 2026].

[52] L. Ye, T. Yamamoto, "Impact of dedicated lanes for connected and autonomous vehicle on traffic flow throughput," *Physica A: Statistical Mechanics and its Applications*, vol. 512, pp. 588–597, Dec, 2018, doi: 10.1016/j.physa.2018.08.083. [Online]. Available: https://doi.org/10.1016/j.physa.2018.08.083.

[53] S. R. Rad, H. Farah, H. Taale, B. v. Arem, S. P. Hoogendoorn, "Design and operation of dedicated lanes for connected and automated vehicles on motorways: A conceptual framework and research agenda," *Transportation Research Part C: Emerging Technologies*, vol. 117, pp. 102664, Aug, 2020, doi: 10.1016/j.trc.2020.102664. [Online]. Available: https://doi.org/10.1016/j.trc.2020.102664.

[54] K. Hartman and L. Shields, “State Policies Promoting Hybrid and Electric Vehicles,” 2021. [Online]. Available: https://www.ncsl.org/research/energy/state-electric-vehicle-incentives-state-chart.aspx. [Accessed: 26 January 2026].

[55] B. Dooley and H. Ueno, “Why Japan Is Holding Back as the World Rushes Toward Electric Cars,” 2021. [Online]. Available: https://www.nytimes.com/2021/03/09/business/electric-cars-japan.html. [Accessed: 26 January 2026].

[56] M. Farrer, “Why Japan’s carmaking heavyweights could be facing an electric shock,” 2021. [Online]. Available: https://www.theguardian.com/environment/2021/mar/18/why-japans-carmaking-heavyweights-could-be-facing-an-electric-shock. [Accessed: 26 January 2026].

[57] IEA, *Japan 2021 Energy Policy Review*, International Energy Agency, 2021. [Online]. Available: https://iea.blob.core.windows.net/assets/3470b395-cfdd-44a9-9184-0537cf069c3d/Japan2021_EnergyPolicyReview.pdf. [Accessed: 26 January 2026].

[58] X. Hu, Z. Yang, J. Sun, Y. Zhang, "Exempting battery electric vehicles from traffic restrictions: Impacts on market and environment under Pigovian taxation," *Transportation Research Part A: Policy and Practice*, vol. 154, pp. 53–91, Dec, 2021, doi: 10.1016/j.tra.2021.09.014. [Online]. Available: https://doi.org/10.1016/j.tra.2021.09.014.

[59] T. L. Sheldon, R. Dua, "Effectiveness of China extquotesingles plug-in electric vehicle subsidy," *Energy Economics*, vol. 88, pp. 104773, May, 2020, doi: 10.1016/j.eneco.2020.104773. [Online]. Available: https://doi.org/10.1016/j.eneco.2020.104773.

[60] T. L. Sheldon, R. Dua, "Measuring the cost-effectiveness of electric vehicle subsidies," *Energy Economics*, vol. 84, pp. 104545, October, 2019, doi: 10.1016/j.eneco.2019.104545. [Online]. Available: https://doi.org/10.1016/j.eneco.2019.104545.

[61] C. M"unzel, P. Pl"otz, F. Sprei, T. Gnann, "How large is the effect of financial incentives on electric vehicle sales? extendash A global review and European analysis," *Energy Economics*, vol. 84, pp. 104493, October, 2019, doi: 10.1016/j.eneco.2019.104493. [Online]. Available: https://doi.org/10.1016/j.eneco.2019.104493.

[62] R. Azarafshar, W. N. Vermeulen, "Electric vehicle incentive policies in Canadian provinces," *Energy Economics*, vol. 91, pp. 104902, September, 2020, doi: 10.1016/j.eneco.2020.104902. [Online]. Available: https://doi.org/10.1016/j.eneco.2020.104902.

[63] CBS, “Hertz to buy 100,000 Tesla cars in push to offer electric vehicles,” 2021. [Online]. Available: https://www.cbsnews.com/news/hertz-100000-tesla-model-3-car-rental/. [Accessed: 26 January 2026].

[64] M. Gorner and L. Paoli, “How global electric car sales defied Covid-19 in 2020,” International Energy Agency, International Energy Agency, 2021. [Online]. Available: https://www.iea.org/commentaries/how-global-electric-car-sales-defied-covid-19-in-2020. [Accessed: 26 January 2026].

[65] Autovista-Group, "How has COVID-19 impacted fleets?," Autovista Group, 2020. [Online]. Available: https://autovistagroup.com/news-and-insights/how-has-covid-19-impacted-fleets. [Accessed: 26 January 2026].

[66] K. J. Schaefer, L. Tuitjer, M. Levin-Keitel, "Transport disrupted extendash Substituting public transport by bike or car under Covid 19," *Transportation Research Part A: Policy and Practice*, vol. 153, pp. 202–217, Nov, 2021, doi: 10.1016/j.tra.2021.09.002. [Online]. Available: https://doi.org/10.1016/j.tra.2021.09.002.

[67] G. Currie, T. Jain, L. Aston, "Evidence of a post-COVID change in travel behaviour extendash Self-reported expectations of commuting in Melbourne," *Transportation Research Part A: Policy and Practice*, vol. 153, pp. 218–234, Nov, 2021, doi: 10.1016/j.tra.2021.09.009. [Online]. Available: https://doi.org/10.1016/j.tra.2021.09.009.

[68] X. Jin and A. Meintz, "Challenges and Opportunities for Transactive Control of Electric Vehicle Supply Equipment: A Reference Guide," National Renewable Energy Laboratory, National Renewable Energy Laboratory, 2015. [Online]. Available: https://www.nrel.gov/docs/fy15osti/64007.pdf. [Accessed: 26 January 2026].

[69] R. Walton, "Xcel's proposed TOU rates could mean big peak demand savings for DER owning customers," Utility Dive, 2019. [Online]. Available: https://www.utilitydive.com/news/xcel-files-residential-tou-rates-in-colorado-following-a-successful-pilot/568642/. [Accessed: 26 January 2026].

[70] D. Thill, "ComEd wins approval to test time-of-use rates starting in 2020," Energy News Network, 2017. [Online]. Available: https://energynews.us/2019/10/21/midwest/comed-wins-approval-to-test-time-of-use-rates-starting-in-2020/. [Accessed: 26 January 2026].

[71] M. Biviji, C. Uckun, G. Bassett, J. Wang, D. Ton, "Patterns of electric vehicle charging with time of use rates: Case studies in California and Portland," pp. 1-5, 2014, doi: 10.1109/ISGT.2014.6816454. [Online]. Available: https://doi.org/10.1109/ISGT.2014.6816454.

[72] A. Masood, J. Hu, A. Xin, A. R. Sayed, G. Yang, "Transactive Energy for Aggregated Electric Vehicles to Reduce System Peak Load Considering Network Constraints," *IEEE Access*, vol. 8, pp. 31519-31529, 2020, doi: 10.1109/ACCESS.2020.2973284. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2973284.

[73] M. Gray, "Analysis and Evaluation of Transactive Energy Control in Active Distribution Systems," *University of Ontario Institute of Technology*, 2016. [Online]. Available: https://ir.library.ontariotechu.ca/handle/10155/738.

[74] S. A. Madkour, "Transactive Energy control of electric energy storage to mitigate the impact of transportation electricification in distribution systems," *University of Ontario Institute of Technology*, 2016. [Online]. Available: https://ir.library.ontariotechu.ca/handle/10155/734.

[75] Y. Wu, Y. Wu, J. M. Guerrero, J. C. Vasquez, "Decentralized transactive energy community in edge grid with positive buildings and interactive electric vehicles," *International Journal of Electrical Power and Energy Systems*, vol. 135, pp. 107510, Feb, 2022, doi: 10.1016/j.ijepes.2021.107510. [Online]. Available: https://doi.org/10.1016/j.ijepes.2021.107510.

[76] S. Behboodi, D. P. Chassin, C. Crawford, N. Djilali, "Electric Vehicle Participation in Transactive Power Systems Using Real-Time Retail Prices," pp. 2400-2407, 2016, doi: 10.1109/HICSS.2016.300. [Online]. Available: https://doi.org/10.1109/HICSS.2016.300.

[77] Z. liu, Q. Wu, M. Shahidehpour, C. Li, S. Huang, W. Wei, "Transactive Real-Time Electric Vehicle Charging Management for Commercial Buildings With PV On-Site Generation," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 4939–4950, Sep, 2019, doi: 10.1109/tsg.2018.2871171. [Online]. Available: https://doi.org/10.1109/tsg.2018.2871171.

[78] J. Zhang, T. Markel, "Charge Management Optimization for Future TOU Rates," *World Electric Vehicle Journal*, vol. 8, pp. 521-530, 06, 2016, doi: 10.3390/wevj8020521. [Online]. Available: https://doi.org/10.3390/wevj8020521.

[79] X. Hai, L. Yin, Z. Jia, Q. Yu, Y. Wang, D. Yao, "Optimizing Capacity Configuration of Photovoltaic and Battery Energy Storage Systems in EV Charging Station based on Time-of-Use Pricing," *IOP Conference Series: Materials Science and Engineering*, vol. 486, pp. 012062, 07, 2019, doi: 10.1088/1757-899X/486/1/012062. [Online]. Available: https://doi.org/10.1088/1757-899X/486/1/012062.

[80] J. Hu, S. You, M. Lind, J. Østergaard, "Coordinated Charging of Electric Vehicles for Congestion Prevention in the Distribution Grid," *IEEE Transactions on Smart Grid*, vol. 5, no. 2, pp. 703-711, 2014, doi: 10.1109/TSG.2013.2279007. [Online]. Available: https://doi.org/10.1109/TSG.2013.2279007.

[81] J. Hu, G. Yang, K. Kok, Y. Xue, H. W. Bindner, "Transactive control: a framework for operating power systems characterized by high penetration of distributed energy resources," *Journal of modern power systems and clean energy*, vol. 5, no. 3, pp. 451-464, 2016, doi: 10.1007/s40565-016-0228-1. [Online]. Available: https://doi.org/10.1007/s40565-016-0228-1.

[82] NUVVE, "V2G Chargers," NUVVE, 2021. [Online]. Available: https://nuvve.com/projects/. [Accessed: 26 January 2026].

[83] NREL, "PREPARING DISTRIBUTION UTILITIES FOR UTILITYSCALE STORAGE AND ELECTRIC VEHICLES," National Renewable Energy Laboratory, 2020. [Online]. Available: https://www.nrel.gov/docs/fy20osti/75973.pdf. [Accessed: 26 January 2026].

[84] E. Schmidt, "The impact of growing electric vehicle adoption on electric utility grids," FLEET CARMA, 2017. [Online]. Available: https://www.fleetcarma.com/impact-growing-electric-vehicle-adoption-electric-utility-grids/. [Accessed: 26 January 2026].

[85] J. Yang, Y. Li, Y. Cao, Y. Tan, C. Rehtanz, "Transactive energy system: a review of cyber-physical infrastructure and optimal scheduling," *IET Generation, Transmission |& Distribution*, vol. 14, no. 2, pp. 173-179, 2020, doi: 10.1049/iet-gtd.2018.6554. [Online]. Available: https://doi.org/10.1049/iet-gtd.2018.6554.

[86] Frost-Sullivan, "Strategic Analysis of Japan's Electric Vehicle Charging Infrastructure V2G and V2H Industry to 2020," *China Weekly News*, NewsRX LLC, pp. 180-, 2014. [Online]. Available: https://store.frost.com/strategic-analysis-of-electric-vehicle-charging-infrastructure-v2g-and-v2h-in-japan.html. [Accessed: 26 January 2026].

[87] H. Turker, S. Bacha, "Smart Charging of Plug-in Electric Vehicles (PEVs) in Residential Areas: Vehicle-to-Home (V2H) and Vehicle-to-Grid (V2G) concepts," *International journal of renewable energy research*, vol. 4, no. 4, pp. 859-871, 2014. [Online]. Available: https://www.ijrer.com/index.php/ijrer/article/view/1711. [Accessed: 26 January 2026].

[88] S. Aznavi, P. Fajri, M. B. Shadmand, A. Khoshkbar-Sadigh, "Peer-to-Peer Operation Strategy of PV Equipped Office Buildings and Charging Stations Considering Electric Vehicle Energy Pricing," *IEEE transactions on industry applications*, vol. 56, no. 5, pp. 5848-5857, 2020, doi: 10.1109/TIA.2020.2990585. [Online]. Available: https://doi.org/10.1109/TIA.2020.2990585.

[89] S. Thakur, B. P. Hayes, J. G. Breslin, "A unified model of peer to peer energy trade and electric vehicle charging using blockchains," pp. 1-6, 2018, doi: 10.1049/cp.2018.1909. [Online]. Available: https://doi.org/10.1049/cp.2018.1909.

[90] E. Myers, “Hope or Only Hype for Residential V2G?,” Smart Electric Power Alliance, 2020. [Online]. Available: https://sepapower.org/knowledge/hope-or-only-hype-for-residential-v2g/. [Accessed: 26 January 2026].

[91] F. Lambert, “Tesla quietly adds bidirectional charging capability for game-changing new features,” Electrek, 2019. [Online]. Available: https://electrek.co/2020/05/19/tesla-bidirectional-charging-ready-game-changing-features/. [Accessed: 26 January 2026].

[92] D. Alfaro, “Is the future of EV charging bidirectional?,” Renewable Energy World, 2020. [Online]. Available: https://www.renewableenergyworld.com/storage/is-the-future-of-ev-charging-bidirectional/. [Accessed: 26 January 2026].

[93] S. Weintraub, “Wallbox Quasar bidirectional home DC charger will turn EVs into a huge Tesla Powerwall,” Electrek, 2020. [Online]. Available: https://electrek.co/2020/01/06/wallbox-quasar-tesla-nissan/. [Accessed: 26 January 2026].

[94] M. Alinejad, O. Rezaei, A. Kazemi, S. Bagheri, "An Optimal Management for Charging and Discharging of Electric Vehicles in an Intelligent Parking Lot Considering Vehicle Owner's Random Behaviors," *Journal of energy storage*, vol. 35, pp. 102245-, 2021, doi: 10.1016/j.est.2021.102245. [Online]. Available: https://doi.org/10.1016/j.est.2021.102245.

[95] L. Zhang, Y. Li, "A Game-Theoretic Approach to Optimal Scheduling of Parking-Lot Electric Vehicle Charging," *IEEE transactions on vehicular technology*, vol. 65, no. 6, pp. 4068-4078, 2016, doi: 10.1109/TVT.2015.2487515. [Online]. Available: https://doi.org/10.1109/TVT.2015.2487515.

[96] L. Zhang, Y. Li, "Optimal Management for Parking-Lot Electric Vehicle Charging by Two-Stage Approximate Dynamic Programming," *IEEE transactions on smart grid*, vol. 8, no. 4, pp. 1722-1730, 2017, doi: 10.1109/TSG.2015.2505298. [Online]. Available: https://doi.org/10.1109/TSG.2015.2505298.

[97] S. M. A. H. Kandil, “Planning of PEVs Parking Lots in Conjunction With Renewable Energy Resources and Battery Energy Storage Systems,” York University, 2015. [Online]. Available: http://hdl.handle.net/10315/32180. [Accessed: 26 January 2026].

[98] S. Hussain, M. A. Ahmed, K. Lee, Y. Kim, "Fuzzy Logic Weight Based Charging Scheme for Optimal Distribution of Charging Power among Electric Vehicles in a Parking Lot," *Energies (Basel)*, vol. 13, no. 12, pp. 3119-, 2020, doi: 10.3390/en13123119. [Online]. Available: https://doi.org/10.3390/en13123119.

[99] S. Powell, E. C. Kara, R. Sevlian, G. V. Cezar, S. Kiliccote, R. Rajagopal, "Controlled workplace charging of electric vehicles: The impact of rate schedules on transformer aging," *Applied energy*, vol. 276, pp. 115352-, 2020, doi: 10.1016/j.apenergy.2020.115352. [Online]. Available: https://doi.org/10.1016/j.apenergy.2020.115352.

[100] A. M. Haider, K. M. Muttaqi, M. H. Haque, "Multistage time-variant electric vehicle load modelling for capturing accurate electric vehicle behaviour and electric vehicle impact on electricity distribution grids," *IET Generation, Transmission |& Distribution*, vol. 9, no. 16, pp. 2705-2716, 2015, doi: 10.1049/iet-gtd.2014.1019. [Online]. Available: https://doi.org/10.1049/iet-gtd.2014.1019.

[101] A. D. Hilshey, P. Rezaei, P. D. H. Hines, J. Frolik, "Electric vehicle charging: Transformer impacts and smart, decentralized solutions," pp. 1-8, July, 2012, doi: 10.1109/PESGM.2012.6345472. [Online]. Available: https://doi.org/10.1109/PESGM.2012.6345472.

[102] M. J. Rutherford, V. Yousefzadeh, "The impact of Electric Vehicle battery charging on distribution transformers," pp. 396-400, Mar, 2011, doi: 10.1109/apec.2011.5744627. [Online]. Available: https://doi.org/10.1109/apec.2011.5744627.

[103] R. B. Bass, N. Zimmerman, "Impacts of Electric Vehicle Charging on Electric Power Distribution Systems," Sep, 2013, doi: 10.15760/trec.145. [Online]. Available: https://doi.org/10.15760/trec.145.

[104] P. Hines, J. Frolik, J. Marshall, P. Rezaei, A. Seier, A. Fuhrmann, J. R. Dowds, and A. Hilshey, "Understanding and Managing the Impacts of Electric Vehicles on Electric Power Distribution Systems," University of Vermont Transportation Research Center, 2014. [Online]. Available: https://www.uvm.edu/sites/default/files/Transportation-Research-Center/Reports/2014/Understanding_and_Managing_the_Impacts_of_Electric_Vehicles_on_Electric_Power_Distribution_Systems.pdf. [Accessed: 26 January 2026].

[105] G. Hallevy, "Unmanned Vehicles - Subordination to Criminal Law under the Modern Concept of Criminal Liability," *Journal of Law, Information & Science*, vol. 21, no. 2, Jan. 2011, doi: 10.5778/JLIS.2011.21.Hallevy.1. [Online]. Available: https://doi.org/10.5778/JLIS.2011.21.Hallevy.1.

[106] R. C. Arkin, "Ethics and Autonomous Systems: Perils and Promises [Point of View]," in *Proceedings of the IEEE*, vol. 104, no. 10, pp. 1779-1781, Oct. 2016, doi: 10.1109/JPROC.2016.2601162. [Online]. Available: https://doi.org/10.1109/JPROC.2016.2601162.

[107] V. Dignum, "Ethics in artificial intelligence: introduction to the special issue," *Ethics and Information Technology*, vol. 20, no. 1, pp. 1-3, Feb. 13, 2018, doi: 10.1007/s10676-018-9450-z. [Online]. Available: https://doi.org/10.1007/s10676-018-9450-z.

[108] N. Adnan, S. Md Nordin, M. A. bin Bahruddin, M. Ali, "How trust can drive forward the user acceptance to the technology? In-vehicle technology for autonomous vehicle," *Transportation Research Part A: Policy and Practice*, vol. 118, pp. 819-836, Dec. 2018. doi: 10.1016/j.tra.2018.10.019. [Online]. Available: https://doi.org/10.1016/j.tra.2018.10.019.

[109] N. Adnan, S. M. Nordin, M. A. b. Bahruddin, "Sustainable Interdependent Networks from Smart Autonomous Vehicle to Intelligent Transportation Networks," in *Sustainable Interdependent Networks II*, *Springer International Publishing*, pp. 121–134, Dec. 12, 2018. doi: 10.1007/978-3-319-98923-5_7. [Online]. Available: https://doi.org/10.1007/978-3-319-98923-5_7.

[110] S. HRYNKO, R. HRYNKO, "Autonomous Car as a Source of Damage: Civil Law Aspect," *University Scientific Notes, Civil Law and Civil Process, Leonid Yuzkov Khmelnytskyi University of Management and Law*, no. 3(71), pp. 91–100, Dec. 27, 2019. doi: 10.37491/unz.71.8. [Online]. Available: https://doi.org/10.37491/unz.71.8.

[111] P. Jha, K. S. Patnaik, "Self-Driving Cars: Role of Machine Learning," in *Handbook of Research on Emerging Trends and Applications of Machine Learning*, New York, NY, USA: IGI Global, Jan. 2020, pp. 490–507. doi: 10.4018/978-1-5225-9643-1.ch023. [Online]. Available: https://doi.org/10.4018/978-1-5225-9643-1.ch023.

[112] D. Peters, K. Vold, D. Robinson, R. A. Calvo, "Responsible AI-Two Frameworks for Ethical Design Practice," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 34-47, Mar. 2020, doi: 10.1109/tts.2020.2974991. [Online]. Available: https://doi.org/10.1109/tts.2020.2974991.

[113] C. Wiedeman, G. Wang, U. Kruger, "Modeling of moral decisions with deep learning," *Visual Computing for Industry, Biomedicine, and Art*, vol. 3, no. 1, pp. 27, Nov. 2020, doi: 10.1186/s42492-020-00063-9. [Online]. Available: https://doi.org/10.1186/s42492-020-00063-9.

[114] M. Chikaraishi, D. Khan, B. Yasuda, A. Fujiwara, "Risk perception and social acceptability of autonomous vehicles: A case study in Hiroshima, Japan," *Transport Policy*, vol. 98, pp. 105-115, Nov. 2020, doi: 10.1016/j.tranpol.2020.05.014. [Online]. Available: https://doi.org/10.1016/j.tranpol.2020.05.014.

[115] International Organization for Standardization, "ISO - Building a responsible AI: How to manage the AI ethics debate," ISO.org. [Online]. Available: https://www.iso.org/artificial-intelligence/responsible-ai-ethicsâĂŒ. [Accessed: 26 January 2026].

[116] A. Jobin, M. Ienca, E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389-399, Sep. 2019, doi: 10.1038/s42256-019-0088-2. [Online]. Available: https://doi.org/10.1038/s42256-019-0088-2.

[117] I. S. Bangroo, "AI-based Predictive Analytic Approaches for safeguarding the Future of Electric/Hybrid Vehicles," *arXiv.org*, Apr. 26, 2023, doi: 10.48550/ARXIV.2304.13841. [Online]. Available: https://doi.org/10.48550/ARXIV.2304.13841.

[118] M. Rauf, L. Kumar, S. A. Zulkifli, A. Jamil, "Aspects of artificial intelligence in future electric vehicle technology for sustainable environmental impact," *Environmental Challenges*, vol. 14, no. 100854, Jan. 2024, doi: 10.1016/j.envc.2024.100854. [Online]. Available: https://doi.org/10.1016/j.envc.2024.100854.

[119] A. Kumar, "Exploring Ethical Considerations in AI-driven Autonomous Vehicles: Balancing Safety and Privacy," *Journal of Artificial Intelligence General science (JAIGS) ISSN:3006-4023*, vol. 2, no. 1, pp. 125-138, Mar. 2024, doi: 10.60087/jaigs.v2i1.p138. [Online]. Available: https://doi.org/10.60087/jaigs.v2i1.p138.

[120] V. Dubljević, G. F. List, J. Milojevich, N. Ajmeri, W. Bauer, M. P. Singh, E. Bardaka, T. Birkland, C. Edwards, R. Mayer, I. Muntean, T. Powers, H. Rakha, V. Ricks, M. S. Samandar, "Toward a Rational and Ethical Sociotechnical System of Autonomous Vehicles: A Novel Application of Multi-Criteria Decision Analysis," *PLOS ONE*, vol. 16, no. 8, Aug. 2021, doi: 10.48550/ARXIV.2102.02928. [Online]. Available: https://doi.org/10.48550/ARXIV.2102.02928.

[121] M. Ryan, "The Future of Transportation: Ethical, Legal, Social and Economic Impacts of Self-driving Vehicles in the Year 2025," *Science and Engineering Ethics*, vol. 26, no. 3, pp. 1185-1208, Sep. 2020, doi: 10.1007/s11948-019-00130-2. [Online]. Available: https://doi.org/10.1007/s11948-019-00130-2.

[122] S. O. Hansson, M. Belin, B. Lundgren, "Self-Driving Vehicles-an Ethical Overview," *Philosophy & Technology*, vol. 34, no. 4, pp. 1383-1408, Aug. 2021, doi: 10.1007/s13347-021-00464-5. [Online]. Available: https://doi.org/10.1007/s13347-021-00464-5.

[123] M. L. Cummings, "What Self-Driving Cars Tell Us About AI Risks," IEEE Spectrum. Accessed: Mar. 28, 2025. [Online.] Available: https://spectrum.ieee.org/self-driving-cars-2662494269. [Accessed: 26 January 2026].

[124] International Organization for Standardization, "ISO 26262: Road vehicles - Functional safety", Standard ISO. [Online.] Available: https://www.iso.org/standard/68383.html. [Accessed: 26 January 2026].

[125] International Organization for Standardization, "ISO 21448: Road vehicles - Safety of the intended functionality", Standard ISO. [Online.] Available: https://www.iso.org/standard/70939.html. [Accessed: 26 January 2026].

[126] A. Giannaros, A. Karras, L. Theodorakopoulos, C. Karras, P. Kranias, N. Schizas, G. Kalogeratos, D. Tsolis, "Autonomous Vehicles: Sophisticated Attacks, Safety Issues, Challenges, Open Topics, Blockchain, and Future Directions," *Journal of Cybersecurity and Privacy*, vol. 3, no. 3, pp. 493-543, Aug. 2023, doi: 10.3390/jcp3030025. [Online]. Available: https://doi.org/10.3390/jcp3030025.

[127] T. Nguyen, T. G. Vu, H. Tran, K. Wong, "Emerging Privacy and Trust Issues for Autonomous Vehicle Systems," in *2022 International Conference on Information Networking (ICOIN)*, IEEE, Jan. 2022, pp. 52-57, doi: 10.1109/ICOIN53446.2022.9687196. [Online]. Available: https://doi.org/10.1109/ICOIN53446.2022.9687196.

[128] S. Krügel, M. Uhl, "The risk ethics of autonomous vehicles: an empirical approach," *Scientific Reports*, vol. 14, no. 1, Jan. 2024, doi: 10.1038/s41598-024-51313-2. [Online]. Available: https://doi.org/10.1038/s41598-024-51313-2.

[129] M. L. Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems," in *Decision making in aviation*, Routledge, 2017, pp. 289-294. doi: 10.2514/6.2004-6313. [Online]. Available: https://doi.org/10.2514/6.2004-6313.

[130] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1-37, Mar. 2014, doi: 10.1145/2523813. [Online]. Available: https://doi.org/10.1145/2523813.

[131] J. Pattinson, H. Chen, S. Basu, "Legal issues in automated vehicles: critically considering the potential role of consent and interactive digital interfaces," *Humanities and Social Sciences Communications*, vol. 7, no. 1, Nov. 2020, doi: 10.1057/s41599-020-00644-2. [Online]. Available: https://doi.org/10.1057/s41599-020-00644-2.

[132] J. Caporal, J. Lim, S. Arrieta-Kenna, and W. O'Neil, "Driving the Future of AV Regulations: Barriers to Large-Scale Development," 2021. [Online]. Available: https://www.csis.org/analysis/driving-future-av-regulations-barriers-large-scale-development. Accessed: 26 January 2026.

[133] J. Reimer, "Debate worth having: will autonomous vehicles take millions of jobs?," 2022. [Online]. Available: https://www.here.com/learn/blog/autonomous-vehicles-jobs. Accessed: 26 January 2026.

[134] P. Kopelias, E. Demiridi, K. Vogiatzis, A. Skabardonis, V. Zafiropoulou, "Connected & autonomous vehicles - Environmental impacts - A review," *Science of The Total Environment*, vol. 712, no. 135237, Apr. 2020, doi: 10.1016/j.scitotenv.2019.135237. [Online]. Available: https://doi.org/10.1016/j.scitotenv.2019.135237.

[135] OpenDP Project Team, *The OpenDP White Paper*. Harvard University, May 2020. Available: https://projects.iq.harvard.edu/files/opendp/files/opendp_white_paper_11may2020.pdf. [Accessed: 26 January 2026].

[136] U.S. National Science and Technology Council, Office of Science and Technology Policy, *National Strategy to Advance Privacy-Preserving Data Sharing and Analytics*. Executive Office of the President of the United States, 2023. Available: https://www.nitrd.gov/pubs/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf. [Accessed: 26 January 2026].

[137] Y. Zhang, L. Chen, W. Li, *et al.*, "Exploring the Tradeoff Between Data Privacy and Utility with a Clinical Analytic Use Case," *BMC Medical Informatics and Decision Making*, vol. 24, Article 123, 2024. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11137882/. [Accessed: 26 January 2026].

[138] D. Kurtz, S. Lee, C. Martínez, *et al.*, "Privacy–Utility Trade-Offs in Genetic Data Sharing and Medical Research," in *Proceedings of the 45th International Conference on Information Systems (ICIS 2024)*. AIS eLibrary, 2024. Available: https://aisel.aisnet.org/icis2024/security/security/8/. [Accessed: 26 January 2026].

[139] N. Li, M. Lyu, D. Su, and W. Yang, Differential Privacy: From Theory to Practice. in Synthesis Lectures on Information Security, Privacy, and Trust. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-031-02350-7. [Online]. Available: https://link.springer.com/10.1007/978-3-031-02350-7.

[140] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," FNT in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2013, doi: 10.1561/0400000042. [Online]. Available: http://www.nowpublishers.com/articles/foundations-and-trends-in-theoretical-computer-science/TCS-042.

[141] C. Dwork, "Differential Privacy," in Automata, Languages and Programming, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12. doi: 10.1007/11787006_1. [Online]. Available: http://link.springer.com/10.1007/11787006_1.

[142] L. Chen, R. Lu, Z. Cao, K. AlHarbi, X. Lin, "MuDA: Multifunctional data aggregation in privacy-preserving smart grid communications," *Peer-to-Peer Networking and Applications*, *Springer Science and Business Media LLC*, vol. 8, no. 5, pp. 777–792, 2014, doi: 10.1007/s12083-014-0292-0. [Online]. Available: http://link.springer.com/10.1007/s12083-014-0292-0.

[143] A. Rial, G. Danezis, M. Kohlweiss, "Privacy-preserving smart metering revisited," *International Journal of Information Security*, *Springer Science and Business Media LLC*, vol. 17, no. 1, pp. 1–31, 2016, doi: 10.1007/s10207-016-0355-8. [Online]. Available: http://link.springer.com/10.1007/s10207-016-0355-8.

[144] G. Giaconi, D. Gunduz, H. V. Poor, "Smart Meter Privacy With Renewable Energy and an Energy Storage Device," *IEEE Transactions on Information Forensics and Security*, *Institute of Electrical and Electronics Engineers (IEEE)*, vol. 13, no. 1, pp. 129–142, 2018, doi: 10.1109/TIFS.2017.2744601. [Online]. Available: https://ieeexplore.ieee.org/document/8016368/.

[145] Y. Son, J. Im, H. Kwon, S. Jeon, M. Lee, "Privacy-Preserving Peer-to-Peer Energy Trading in Blockchain-Enabled Smart Grids Using Functional Encryption," *Energies*, *MDPI AG*, vol. 13, no. 6, pp. 1321, 2020, doi: 10.3390/en13061321. [Online]. Available: https://www.mdpi.com/1996-1073/13/6/1321.

[146] R. Dong, L. J. Ratliff, A. A. Cárdenas, H. Ohlsson, S. S. Sastry, "Quantifying the Utility–Privacy Tradeoff in the Internet of Things," *ACM Transactions on Cyber-Physical Systems*, *Association for Computing Machinery (ACM)*, vol. 2, no. 2, pp. 1–28, 2018, doi: 10.1145/3185511. [Online]. Available: https://dl.acm.org/doi/10.1145/3185511.

[147] H. Cao, S. Liu, R. Zhao, X. Xiong, "IFed: A novel federated learning framework for local differential privacy in Power Internet of Things," *International Journal of Distributed Sensor Networks*, *SAGE Publications*, vol. 16, no. 5, pp. 155014772091969, 2020, doi: 10.1177/1550147720919698. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1550147720919698.

[148] M. Islam, M. H. Rehmani, L. Gao, J. Chen, "An Optimized Privacy-Utility Tradeoff Framework for Differentially Private Data Sharing in Blockchain-Based Internet of Things," *IEEE Internet of Things Journal*, *Institute of Electrical and Electronics Engineers (IEEE)*, vol. 12, no. 7, pp. 7778–7792, 2025, doi: 10.1109/JIOT.2025.3530247. [Online]. Available: https://ieeexplore.ieee.org/document/10843378.

[149] S. K. Venkatachary, A. Alagappan, L. J. B. Andrews, "Cybersecurity challenges in energy sector (virtual power plants) - can edge computing principles be applied to enhance security?," *Energy Informatics*, *Springer Science and Business Media LLC*, vol. 4, no. 1, 2021, doi: 10.1186/s42162-021-00139-7. [Online]. Available: https://energyinformatics.springeropen.com/articles/10.1186/s42162-021-00139-7.

[150] A. Tiwari, H. Farag, "Analysis and Modeling of Value Creation Opportunities and Governing Factors for Electric Vehicle Proliferation," *Energies*, *MDPI AG*, vol. 16, no. 1, pp. 438, 2022, doi: 10.3390/en16010438. [Online]. Available: https://www.mdpi.com/1996-1073/16/1/438.

[151] A. Tiwari, H. E. Z. Farag, "Responsible AI Framework for Autonomous Vehicles: Addressing Bias and Fairness Risks," *IEEE Access*, *Institute of Electrical and Electronics Engineers (IEEE)*, vol. 13, pp. 58800–58822, 2025, doi: 10.1109/ACCESS.2025.3556781. [Online]. Available: https://ieeexplore.ieee.org/document/10947002.

[152] A. K. Hildebrandt, E. Schömer, A. Hildebrandt, "Metric Differential Privacy on the Special Orthogonal Group SO(3)," *Journal of Cybersecurity and Privacy*, *MDPI AG*, vol. 5, no. 3, pp. 57, 2025, doi: 10.3390/jcp5030057. [Online]. Available: https://www.mdpi.com/2624-800X/5/3/57.

[153] M. Lange, P. Guerra-Balboa, J. Parra-Arnau, T. Strufe, "Balancing Privacy and Utility in Correlated Data: A Study of Bayesian Differential Privacy," *Proceedings of the VLDB Endowment*, *Association for Computing Machinery (ACM)*, vol. 18, no. 11, pp. 4090–4103, 2025, doi: 10.14778/3749646.374967. [Online]. Available: https://dl.acm.org/doi/10.14778/3749646.3749679.

[154] Y. Zhang, Y. Zhu, S. Wang, X. Huang, "Mean Estimation of Numerical Data Under (ε, δ) -Utility-Optimized Local Differential Privacy," *IEEE Transactions on Information Forensics and Security*, *Institute of Electrical and Electronics Engineers (IEEE)*, vol. 19, pp. 9656–9669, 2024, doi: 10.1109/TIFS.2024.3478823. [Online]. Available: https://ieeexplore.ieee.org/document/10714480.

[155] X. Li, S. Dong, A. Milani Fard, "Enhancing User Experience with Visual Controls for Local Differential Privacy," *Journal of Cybersecurity and Privacy*, *MDPI AG*, vol. 5, no. 3, pp. 36, 2025, doi: 10.3390/jcp5030036. [Online]. Available: https://www.mdpi.com/2624-800X/5/3/36.

[156] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, D. Megias, "Individual Differential Privacy: A Utility-Preserving Formulation of Differential Privacy Guarantees," *IEEE Transactions on Information Forensics and Security*, *Institute of Electrical and Electronics Engineers (IEEE)*, vol. 12, no. 6, pp. 1418–1429, 2017, doi: 10.1109/TIFS.2017.2663337. [Online]. Available: https://ieeexplore.ieee.org/document/7839941.

[157] Q. Razi, S. Datta, V. Hassija, G. Chalapathi, B. Sikdar, "Privacy Utility Tradeoff Between PETs: Differential Privacy and Synthetic Data," *IEEE Transactions on Computational Social Systems*, *Institute of Electrical and Electronics Engineers (IEEE)*, vol. 12, no. 2, pp. 473–484, 2025, doi: 10.1109/TCSS.2024.3479317. [Online]. Available: https://ieeexplore.ieee.org/document/10753017.

[158] A. Amiri, "Maximizing data utility while preserving privacy through database fragmentation," *Expert Systems with Applications*, *Elsevier BV*, vol. 273, pp. 126873, 2025, doi: 10.1016/j.eswa.2025.126873. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417425004956.

[159] M. Chamikara, S. I. Jang, I. Oppermann, D. Liu, M. Roberto, S. Ruj, A. Pal, M. Mohammady, S. Camtepe, S. Young, C. Dorrian, N. David, "Towards Usability of Data with Privacy: A Unified Framework for Privacy-Preserving Data Sharing with High Utility," *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security*, *ACM*, pp. 790–806, 2025, doi: 10.1145/3708821.3736187. [Online]. Available: https://dl.acm.org/doi/10.1145/3708821.3736187.

[160] CERB, "Zero-Emission Vehicle Program," California Air Resource Board, 2022. [Online]. Available: https://ww2.arb.ca.gov/our-work/programs/zero-emission-vehicle-program. [Accessed: 26 January 2026].

[161] ICCT, "CHINA - LIGHT-DUTY - NEV," The International Council on Clean Transportation, 2022. [Online]. Available: https://www.transportpolicy.net/standard/china-light-duty-nev/. [Accessed: 26 January 2026].

[162] M. Abdullah, C. Dias, D. Muley, M. Shahin, "Exploring the impacts of COVID-19 on travel behavior and mode preferences," *Transportation Research Interdisciplinary Perspectives*, vol. 8, pp. 100255, November, 2020, doi: 10.1016/j.trip.2020.100255. [Online]. Available: https://doi.org/10.1016/j.trip.2020.100255.

[163] K. Heineke, P. Kampshoff, T. Möller, and T. Wu, "From no mobility to future mobility: Where COVID-19 has accelerated change," McKinsey and Company, 2020. [Online]. Available: https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/from-no-mobility-to-future-mobility-where-covid-19-has-accelerated-change. [Accessed: 26 January 2026].

[164] S. Hausler, K. Heineke, R. Hensley, T. Möller, D. Schwedhelm, and P. Shen, "The impact of COVID-19 on future mobility solutions," McKinsey and Company, 2020. [Online]. Available: https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/the-impact-of-covid-19-on-future-mobility-solutions. [Accessed: 26 January 2026].

[165] X. Yang, "Multi-Objective Optimization," pp. 197–211, 2014, doi: 10.1016/b978-0-12-416743-8.00014-2. [Online]. Available: https://doi.org/10.1016/b978-0-12-416743-8.00014-2.

[166] BloombergNEF, *Electric Vehicle Outlook 2021*, Bloomberg Finance L.P., 2021. [Online]. Available: https://about.bnef.com/electric-vehicle-outlook/. [Accessed: 26 January 2026].

[167] Solar-Reviews, "How much electricity prices increase per year in the U.S.," Solar Reviews, 2021. [Online]. Available: https://www.solarreviews.com/blog/average-electricity-cost-increase-per-year. [Accessed: 26 January 2026].

[168] B. Preston, "CR research shows that EVs cost less to maintain than gasoline-powered vehicles," Consumer Reports, 2020. [Online]. Available: https://www.consumerreports.org/car-repair-maintenance/pay-less-for-vehicle-maintenance-with-an-ev/. [Accessed: 26 January 2026].

[169] EE-News, "The future was supposed to be electric. Is it still?," 2020. [Online]. Available: https://www.eenews.net/stories/1062876693?utm_term=0_173e047b1f-094d8adbde-246816833. [Accessed: 26 January 2026].

[170] R. Walton, "'An enormous lift': Biden's goal of 50% EV sales by 2030 will test supply chains, utilities, experts say," Utility Dive, 2021. [Online]. Available: https://www.utilitydive.com/news/an-enormous-lift-bidens-goal-of-50-ev-sales-by-2030-will-test-supply-c/604696/. [Accessed: 26 January 2026].

[171] M. Carlier, "U.S. car sales from 1951 to 2021," Statistica, 2022. [Online]. Available: https://www.statista.com/statistics/199974/us-car-sales-since-1951/. [Accessed: 26 January 2026].

[172] CEIC-Data, "Norway Number of Registered Vehicles," CEIC, 2021. [Online]. Available: https://www.ceicdata.com/en/indicator/norway/number-of-registered-vehicles. [Accessed: 26 January 2026].

[173] Statistica-Research-Department, "Number of registered passenger cars in Sweden," Statistica, 2022. [Online]. Available: https://www.statista.com/statistics/732187/number-of-registered-passenger-cars-in-sweden-monthly/. [Accessed: 26 January 2026].

[174] Country-Economy-Team, "Netherlands - New motor vehicle registrations," Country Economy, 2022. [Online]. Available: https://countryeconomy.com/business/car-registrations/netherlands. [Accessed: 26 January 2026].

[175] M. Carlier, "Average price (including tax) of passenger cars in Europe," Statistica, 2022. [Online]. Available: https://www.statista.com/statistics/425095/eu-car-sales-average-prices-in-by-country/. [Accessed: 26 January 2026].

[176] Tesla-Team, "Vehicle Incentives," Tesla, 2022. [Online]. Available: https://www.tesla.com/en_ie/support/incentives. [Accessed: 26 January 2026].

[177] TNTA-Team, "Cars and other vehicles," The-Norwegian-Tax-Administration, 2022. [Online]. Available: https://www.skatteetaten.no/en/person/duties/cars-and-other-vehicles/. [Accessed: 26 January 2026].

[178] Transport-Styrelsen-Team, "Vehicle Tax," Transport Styrelsen, 2022. [Online]. Available: https://www.transportstyrelsen.se/en/road/vehicles/vehicle-tax/. [Accessed: 26 January 2026].

[179] Netherlands-RVO-Team, "Motor vehicle tax (mrb)," Netherlands Enterprise Agency, RVO, 2022. [Online]. Available: https://business.gov.nl/regulation/motor-vehicle-tax/. [Accessed: 26 January 2026].

[180] Norsk elbilforening, "Norwegian EV policy," ELBIL, 2022. [Online]. Available: https://elbil.no/english/norwegian-ev-policy/. [Accessed: 26 January 2026].

[181] N. Sönnichsen, "Number of fuel stations," Statistica, 2022. [Online]. Available: https://www.statista.com/statistics/658000/number-of-petrol-stations-in-the-netherlands/. [Accessed: 26 January 2026].

[182] M. Carlier, “Number of electric car charging stations,” Statistica, 2022. [Online]. Available: https://www.statista.com/statistics/696548/number-of-electric-car-charging-stations-in-norway-by-type/. [Accessed: 26 January 2026].

[183] MER-Team, “A Look Into Sweden’s EV Charging Infrastructure,” MER, 2022. [Online]. Available: https://uk.mer.eco/news/sweden-ev-charging-infrastructure/. [Accessed: 26 January 2026].

[184] EV-Monitor-Team, “Electric Vehicles Statistics in the Netherlands,” Netherlands Enterprise Agency, 2022. [Online]. Available: https://www.rvo.nl/sites/default/files/2022-07/Statistics-electric-vehicles-and-charging-in-the-%20Netherlands-up-to-and-including-June-2022_0.pdf. [Accessed: 26 January 2026].

[185] World-Data-Atlas, “Sweden - Passenger car sales,” KNOEMA, 2022. [Online]. Available: https://knoema.com/atlas/Sweden/topics/Transportation/Motor-Vehicle-Sales/Car-sales#. [Accessed: 26 January 2026].

[186] CBS-Team, “Motor vehicles; type, age class, 1 January, 2000-2022,” CBS, 2022. [Online]. Available: https://www.cbs.nl/en-gb/figures/detail/82044ENG. [Accessed: 26 January 2026].

[187] Global-EV-Outlook, “Trends in electric light-duty vehicles,” IEA, 2022. [Online]. Available: https://www.iea.org/reports/global-ev-outlook-2022/trends-in-electric-light-duty-vehicles. [Accessed: 26 January 2026].

[188] GPP-Team, “Retail energy price data,” Global Petrol Prices, 2022. [Online]. Available: https://www.globalpetrolprices.com. [Accessed: 26 January 2026].

[189] Country-Economy-Team, “Sweden - Household electricity prices,” Country Economy, 2022. [Online]. Available: https://countryeconomy.com/energy-and-environment/electricity-price-household/sweden. [Accessed: 26 January 2026].

[190] T. Elli, G. Colombo, B. Gobbo, others, "Data, Algorithms and Otherness. The Erasure of the Other," *Design meets alterity: case studies, project experiences, communication criticism*, Torrossa, 2024, pp. 62-80. [Online]. Available: https://www.torrossa.com/en/resources/an/5763291#page=63.

[191] N. Shahbazi, Y. Lin, A. Asudeh, H. V. Jagadish, "Representation Bias in Data: A Survey on Identification and Resolution Techniques," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1-39, Jul. 13, 2023, doi: 10.1145/3588433. [Online]. Available: https://doi.org/10.1145/3588433.

[192] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1-35, Jul. 13, 2021, doi: 10.1145/3457607. [Online]. Available: https://doi.org/10.1145/3457607.

[193] H. Suresh, J. Guttag, "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, ACM, pp. 1–9, Oct. 05, 2021, doi: 10.1145/3465416.3483305. [Online]. Available: https://doi.org/10.1145/3465416.3483305.

[194] Y. Chen, E. W. Clayton, L. L. Novak, S. Anders, B. Malin, "Human-Centered Design to Address Biases in Artificial Intelligence," *Journal of Medical Internet Research*, vol. 25, no. e43251, Mar. 24, 2023, doi: 10.2196/43251. [Online]. Available: https://doi.org/10.2196/43251.

[195] D. Roselli, J. Matthews, N. Talagala, "Managing Bias in AI," presented at the *Companion Proceedings of The 2019 World Wide Web Conference*, ACM, pp. 539–544, May 13, 2019, doi: 10.1145/3308560.3317590. [Online]. Available: https://doi.org/10.1145/3308560.3317590.

[196] V. V. Hernandes, C. Barbas, D. Dudzik, "A review of blood sample handling and pre-processing for metabolomics studies," *ELECTROPHORESIS*, Wiley, vol. 38, no. 18, pp. 2232-2241, Jun. 08, 2017, doi: 10.1002/elps.201700086. [Online]. Available: https://doi.org/10.1002/elps.201700086.

[197] J. Christensen, N. Bajaj, M. Gosada, K. D. Borne, "No Title," in *Data-Centric Machine Learning with Python: The ultimate guide to engineering and deploying high-quality models based on good data*, Packt Publishing, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10460906.

[198] V. Abhishek, K. Hosanagar, P. S. Fader, "Aggregation Bias in Sponsored Search Data: The Curse and the Cure," *Marketing Science*, vol. 34, no. 1, pp. 59-77, Jan. 2015, doi: 10.1287/mksc.2014.0884. [Online]. Available: https://doi.org/10.1287/mksc.2014.0884.

[199] B. J. Casad and J. E. Luebering, "Confirmation Bias," Encyclopedia Britannica, 30 Jul 2024. [Online]. Available: https://www.britannica.com/science/confirmation-bias. [Accessed: 26 January 2026].

[200] Alan Turing Institute, "Deployment Bias," 2024. [Online]. Available: https://alan-turing-institute.github.io/rrp-selfassessment/bias/deployment-biases.html. [Accessed: 26 January 2026].

[201] Domino Data Lab, "Model Drift" Domino.ai. [Online]. Available: https://domino.ai/data-science-dictionary/model-driftâĂŒ. [Accessed: 26 January 2026].

[202] K. Rajashekar, S. Paul, S. Karmakar, and S. Sidhanta, "Minimizing Data Retrieval Delay in Edge Computing," in *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer Nature Switzerland*, pp. 63–85, 2024. doi: 10.1007/978-3-031-63992-0_5. [Online]. Available: https://doi.org/10.1007/978-3-031-63992-0_5.

[203] Y. Wang, M. M. Khalili, X. Zhang, "Towards Fair Representation Learning in Knowledge Graph with Stable Adversarial Debiasing," *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, IEEE, pp. 901-909, 2022, doi: 10.1109/icdmw58026.2022.00119. [Online]. Available: https://doi.org/10.1109/icdmw58026.2022.00119.

[204] M. Li, H. Sun, Y. Huang, H. Chen, "Shapley value: from cooperative game to explainable artificial intelligence," *Autonomous Intelligent Systems*, Springer Science and Business Media LLC, vol. 4, no. 1, 2024, doi: 10.1007/s43684-023-00060-8. [Online]. Available: https://doi.org/10.1007/s43684-023-00060-8.

[205] M. Franzese, A. Iuliano, "Hidden Markov Models," *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, pp. 753-762, 2019, doi: 10.1016/B978-0-12-809633-8.20488-3. [Online]. Available: https://doi.org/10.1016/B978-0-12-809633-8.20488-3.

[206] M. J. Kusner, J. Loftus, C. Russell, R. Silva, "Counterfactual Fairness," presented at the *Advances in Neural Information Processing Systems*, vol. 30, 2017, doi: 10.48550/arXiv.1703.06856. [Online]. Available: https://arxiv.org/abs/1703.06856.

[207] H. Liu, Y. Ong, X. Shen, J. Cai, "When Gaussian Process Meets Big Data: A Review of Scalable GPs," *IEEE Transactions on Neural Networks and Learning Systems*, IEEE, vol. 31, no. 11, pp. 4405-4423, 2020, doi: 10.1109/TNNLS.2019.2957109. [Online]. Available: https://doi.org/10.1109/TNNLS.2019.2957109.

[208] M. Alzahrani, Q. Wang, W. Liao, X. Chen, W. Yu, "Survey on Multi-Task Learning in Smart Transportation," *IEEE Access*, IEEE, vol. 12, pp. 17023-17044, 2024, doi: 10.1109/ACCESS.2024.3355034. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3355034.

[209] P. Iversen, S. Witzke, K. Baum, B. Y. Renard, "Identifying drivers of predictive uncertainty using variance feature attribution," *OpenReview*, 2023. [Online]. Available: https://openreview.net/forum?id=XKgTNCLqW9. [Accessed: 26 January 2026].

[210] A. G. Wilson, Z. Hu, R. Salakhutdinov, E. P. Xing, "Deep kernel learning," presented at the *Proceedings of the 19th International Conference on Artificial intelligence and statistics*, PMLR 51:370-378, 2016. [Online]. Available: https://proceedings.mlr.press/v51/wilson16.html. [Accessed: 26 January 2026].

[211] D. Ribeiro, L. M. Matos, G. Moreira, A. Pilastri, P. Cortez, "Isolation Forests and Deep Autoencoders for Industrial Screw Tightening Anomaly Detection," *Computers*, MDPI AG, vol. 11, no. 4, pp. 54, Apr. 08, 2022, doi: 10.3390/computers11040054. [Online]. Available: https://doi.org/10.3390/computers11040054.

[212] S. Moreno-Ãlvarez, M. E. Paoletti, A. J. Sanchez-Fernandez, J. A. Rico-Gallego, L. Han, J. M. Haut, "Federated learning meets remote sensing," *Expert Systems with Applications*, Elsevier BV, vol. 255, no. 124583, Dec. 2024. doi: 10.1016/j.eswa.2024.124583. [Online]. Available: https://doi.org/10.1016/j.eswa.2024.124583.

[213] J. Qian, "Sampling," in *International Encyclopedia of Education*, Elsevier, 2010, pp. 390-395. [Online]. Available: https://www.ets.org/research/policy_research_reports/publications/chapter/2010/ikye.html. [Accessed: 26 January 2026].

[214] P. A. Rogerson, "Spatial Sampling," in *Encyclopedia of Social Measurement*, Elsevier, 2005, pp. 633-638, doi: 10.1016/B978-0-08-044894-7.00294-3. [Online]. Available: https://doi.org/10.1016/B978-0-08-044894-7.00294-3.

[215] B. H. Zhang, B. Lemoine, M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," presented at the *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ACM, Dec. 27, 2018, doi: 10.1145/3278721.3278779. [Online]. Available: https://doi.org/10.1145/3278721.3278779.

[216] J. Murel, "What is regularization?," 2023. [Online]. Available: https://www.ibm.com/topics/regularization. [Accessed: 26 January 2026].

[217] R. González-Sendino, E. Serrano, J. Bajo, "Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making," *Future Generation Computer Systems*, vol. 155, pp. 384-401, Jun. 2024, doi: 10.1016/j.future.2024.02.023. [Online]. Available: https://doi.org/10.1016/j.future.2024.02.023.

[218] M. Hort, Z. Chen, J. M. Zhang, M. Harman, F. Sarro, "Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey," *ACM Journal on Responsible Computing*, vol. 1, no. 2, pp. 1-52, Jun. 2024, doi: 10.1145/3631326. [Online]. Available: https://doi.org/10.1145/3631326.

[219] L. H. Nazer, R. Zatarah, S. Waldrip, J. X. C. Ke, M. Moukheiber, A. K. Khanna, R. S. Hicklen, L. Moukheiber, D. Moukheiber, H. Ma, P. Mathur, "Bias in artificial intelligence algorithms and recommendations for mitigation," *PLOS Digital Health*, vol. 2, no. 6, pp. e0000278, Jun. 2023, doi: 10.1371/journal.pdig.0000278. [Online]. Available: https://doi.org/10.1371/journal.pdig.0000278.

[220] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, Jul. 2019, doi: 10.1147/jrd.2019.2942287. [Online]. Available: https://doi.org/10.1147/jrd.2019.2942287.

[221] R. Fu, Y. Huang, P. V. Singh, "AI and Algorithmic Bias: Source, Detection, Mitigation and Implications," *SSRN Electronic Journal*, Elsevier BV, vol. 7, pp. 39–63, 2020, doi: 10.2139/ssrn.3681517. [Online]. Available: https://doi.org/10.2139/ssrn.3681517.

[222] T. Henzinger, M. Karimi, K. Kueffner, K. Mallik, "Runtime Monitoring of Dynamic Fairness Properties," in *2023 ACM Conference on Fairness, Accountability, and Transparency*, ACM, pp. 604-614, Jun. 12, 2023, doi: 10.1145/3593013.3594028. [Online]. Available: https://doi.org/10.1145/3593013.3594028.

[223] N. Churamani, O. Kara, H. Gunes, "Domain-Incremental Continual Learning for Mitigating Bias in Facial Expression and Action Unit Recognition," *IEEE Transactions on Affective Computing*, *Institute of Electrical and Electronics Engineers (IEEE)*, vol. 14, no. 4, pp. 3191-3206, Oct. 01, 2023, doi: 10.1109/taffc.2022.3181033. [Online]. Available: https://doi.org/10.1109/taffc.2022.3181033.

[224] T. Zhao, W. Zhang, H. Zhao, Z. Jin, "A Reinforcement Learning-Based Framework for the Generation and Evolution of Adaptation Rules," *2017 IEEE International Conference on Autonomic Computing (ICAC)*, IEEE, pp. 103-112, Jul. 2017, doi: 10.1109/icac.2017.47. [Online]. Available: https://doi.org/10.1109/icac.2017.47.

[225] K. Mukherjee, A. Ray, T. Wettergren, S. Gupta, S. Phoha, "Real-time adaptation of decision thresholds in sensor networks for detection of moving targets," *Automatica*, Elsevier BV, vol. 47, no. 1, pp. 185-191, Jan. 2011, doi: 10.1016/j.automatica.2010.10.031. [Online]. Available: https://doi.org/10.1016/j.automatica.2010.10.031.

[226] R. Deokar, P. Nanjundan, S. N. Mohanty, "Transparency in Translation: A Deep Dive into Explainable AI Techniques for Bias Mitigation," *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, IEEE, pp. 1-6, Jul. 26, 2024, doi: 10.1109/apcit62007.2024.10673712. [Online]. Available: https://doi.org/10.1109/apcit62007.2024.10673712.

[227] Y. Djebrouni, N. Benarba, O. Touat, P. De Rosa, S. Bouchenak, A. Bonifati, P. Felber, V. Marangozova, V. Schiavoni, "Bias Mitigation in Federated Learning for Edge Computing," in *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Association for Computing Machinery (ACM), vol. 7, no. 4, pp. 1-35, Dec. 19, 2023, doi: 10.1145/3631455. [Online]. Available: https://doi.org/10.1145/3631455.

[228] L. Zhang, X. Gao, "Transfer Adaptation Learning: A Decade Survey," *IEEE Transactions on Neural Networks and Learning Systems*, IEEE, vol. 35, no. 1, pp. 23-44, Jan. 2024, doi: 10.1109/tnnls.2022.3183326. [Online]. Available: https://doi.org/10.1109/tnnls.2022.3183326.

[229] Z. Zhao, F. Zhou, K. Xu, Z. Zeng, C. Guan, S. K. Zhou, "LE-UDA: Label-Efficient Unsupervised Domain Adaptation for Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, IEEE, vol. 42, no. 3, pp. 633-646, Mar. 2023, doi: 10.1109/tmi.2022.3214766. [Online]. Available: https://doi.org/10.1109/tmi.2022.3214766.

[230] T. Wang, X. Zhang, L. Yuan, J. Feng, "Few-shot adaptive faster r-cnn," presented at the *Proc. IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7173-7182. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Wang_Few-Shot_Adaptive_Faster_R-CNN_CVPR_2019_paper.html. [Accessed: 26 January 2026].

[231] I. Prapas, B. Derakhshan, A. R. Mahdiraji, V. Markl, "Continuous Training and Deployment of Deep Learning Models," *Datenbank-Spektrum*, *Springer Science and Business Media LLC*, vol. 21, no. 3, pp. 203-212, Nov. 2021, doi: 10.1007/s13222-021-00386-8. [Online]. Available: https://doi.org/10.1007/s13222-021-00386-8.

[232] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, F. Wang, "Generative adversarial networks: introduction and outlook," *IEEE/CAA Journal of Automatica Sinica*, IEEE, vol. 4, no. 4, pp. 588-598, 2017, doi: 10.1109/jas.2017.7510583. [Online]. Available: https://doi.org/10.1109/jas.2017.7510583.

[233] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, O. Winther, "Ladder variational autoencoders," in *Proc. Advances in neural information processing systems*, vol. 29, 2016. [Online]. Available: https://papers.nips.cc/paper_files/paper/2016/hash/6ae07dcb33ec3b7c814df797cbda0f87-Abstract.html. [Accessed: 26 January 2026].

[234] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, T. Darrell, "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning," presented at *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 2636-2645, doi: 10.1109/CVPR42600.2020.00271. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00271.

[235] E. Ferrara, "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies," *Sci*, MDPI AG, vol. 6, no. 1, pp. 3, Dec. 26, 2023, doi: 10.3390/sci6010003. [Online]. Available: https://doi.org/10.3390/sci6010003.

[236] ISO/IEC, *Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model*, ISO/IEC 25012:2008, 1st ed., International Organization for Standardization (ISO), Geneva, Switzerland, Dec. 2008. [Online]. Available: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:25012:ed-1:v1:en. Accessed: Jan. 26, 2026.

[237] W. Elouataoui, I. E. Alaoui, S. E. Mendili and Y. Gahi. "An End-to-End Big Data Deduplication Framework based on Online Continuous Learning". International Journal of Advanced Computer Science and Applications (IJACSA) 13.9 (2022), doi: 10.14569/IJACSA.2022.013093. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2022.0130933.

[238] C. E. Shannon, "A mathematical theory of communication," in The Bell System Technical Journal, vol. 27, no. 3, pp. 379-423, July 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x. [Online]. Available: http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[239] P. Shi; Y. Cui; K. Xu.; M. Zhang; L. Ding. Data Consistency Theory and Case Study for Scientific Big Data. Information 2019, 10, 137, doi: 10.3390/info10040137. [Online]. Available: https://doi.org/10.3390/info10040137.

[240] UK Power Networks, "SmartMeter Energy Consumption Data in London Households," London Datastore. Accessed Oct 26, 2025. [Online]. Available: https://data.london.gov.uk/dataset/smartmeter-energy-consumption-data-in-london-households-vqm0d/.

[241] Z. J. Lee, T. Li, and S. H. Low, "ACN-Data: Analysis and Applications of an Open EV Charging Dataset," in Proceedings of the Tenth ACM International Conference on Future Energy Systems, Phoenix AZ USA: ACM, Jun. 2019, pp. 139–149. doi: 10.1145/3307772.3328313. [Online]. Available: https://dl.acm.org/doi/10.1145/3307772.3328313.

[242] R. Shokri, M. Stronati, C. Song and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2017, pp. 3-18, doi: 10.1109/SP.2017.41. [Online]. Available: https://dl.acm.org/doi/10.1109/SP.2017.41.

List of Thesis Publications

[1] A. Tiwari, H. Farag, "Analysis and Modeling of Value Creation Opportunities and Governing Factors for Electric Vehicle Proliferation," *Energies*, *MDPI AG*, vol. 16, no. 1, pp. 438, 2022, doi: 10.3390/en16010438. Available: https://www.mdpi.com/1996-1073/16/1/438.

[2] A. Tiwari, H. E. Z. Farag, "Responsible AI Framework for Autonomous Vehicles: Addressing Bias and Fairness Risks," *IEEE Access*, *Institute of Electrical and Electronics Engineers (IEEE)*, vol. 13, pp. 58800–58822, 2025, doi: 10.1109/ACCESS.2025.3556781. Available: https://ieeexplore.ieee.org/document/10947002

[3] A. Tiwari, H. E. Z. Farag, "A Responsible Generative Artificial Intelligence based Multi-Agent Framework for Preserving Data Utility and Privacy," *Artificial Intelligence*, *MDPI AG*, vol. 7, no. 1, 2025, doi: 10.3390/ai7010001. Available: https://www.mdpi.com/2673-2688/7/1/1.