Conférence Européenne des Directeurs des Routes

Conference of European Directors of Roads

Trusted Integrity and Authenticity for Road Applications (TIARA)

Draft report on connected vehicle deanonymisation research review and impact study

Draft report

This is the first part of the deliverable.
Please address your comments to
sven.maerivoet@tmleuven.be

Deliverable D8

Version 0.2

Date 11/10/2024

CEDR Call 2022 Data: Integrity, Authenticity, and Non-Repudiation integrated in Trust Models for C-ITS applications

Trusted Integrity and Authenticity for Road Applications (TIARA)

D8 Draft report on connected vehicle deanonymisation research review and impact study

Due date of deliverable: 30/06/2024 (Sections 3 and 4)

Due date of deliverable: 30/11/2024 (Section 5)

Due date of deliverable: 30/03/2025 (Section 6)

Due date of deliverable: 31/05/2025 (Section 7)

Actual submission date: 04/07/2024

Start date of project:	End date of project:
22 November 2023	30 June 2025

Author(s): Sven Maerivoet, Bart Ons

Table of contents

List of figures

List of tables

1 Executive summary

Chapter 2 introduces the purpose and scope of the Trusted Integrity and Authenticity for Road Applications (TIARA) project, which seeks to enhance trust in Cooperative Intelligent Transport Systems (C-ITS) by addressing privacy and re-identification risks. The primary goal is to provide practical guidance on ensuring secure and privacy-preserving data infrastructures for road applications. As connected vehicles proliferate, the threat of unauthorszed data access and vehicle re-identification grows. This chapter establishes the framework for the subsequent research on mitigating these privacy risks, focusing on legal and technical perspectives, including trust models, encryption, and regulatory standards. It also provides an extensive list of terminology and nomenclature which is necessary to understand in order to correctly grasp the context of the remaining discussions.

Chapter 3 delves into the current state of connected vehicle deanonymisation research. It explores various re-identification methods used to infer personal information from anonymised vehicle data. Methods such as analysing vehicle GPS traces, behavioural patterns, and temporal data are highlighted as potential risks to individual privacy, despite existing anonymisation efforts. To mitigate these risks, several emerging solutions are presented, including data minimisation, pseudonymisation, and encryption techniques. The chapter also explores advanced privacy-preserving methods such as differential privacy, secure multi-party computation, and federated learning, emphasising the importance of evolving these measures as vehicle technologies advance.

Chapter 4 conducts an impact study of C-ITS data, analysing the potential leakage of sensitive information from Vehicle-to-Everything (V2X) communications. The chapter examines how vehicle data, such as location, speed, and heading, can be correlated with personal data, enabling inferences about individual behaviour, health conditions, and social connections. The impact on data subjects is substantial, as the ability to re-identify individuals poses significant privacy threats. Behavioural profiling and predictive analytics can further heighten these risks. The chapter concludes by underscoring the need for robust legal frameworks, dynamic consent mechanisms, and ongoing enhancements to anonymisation techniques to safeguard user privacy in the rapidly evolving connected vehicle ecosystem.

2 Introduction

2.1 About TIARA

The objective of the TIARA project (*Trusted Integrity and Authenticity for Road Applications*) is to provide the National Road Authorities (NRAs) with an increased understanding of what is required to achieve a trustworthy and secure data infrastructure. The availability of data has allowed road users and NRAs to benefit from new business models. To deliver these benefits, the data infrastructure must be trustworthy and trusted, i.e. secure, with assurances that it is managed to achieve privacy for all stakeholders.

As more C-ITS services develop in Europe, and road users access and share more C-ITS data through open border countries, NRAs will need to ensure greater interoperability through common approaches to connected systems. Data trust is therefore paramount.

CEDR is undertaking a series of projects to research how NRAs can maintain and share the digital road infrastructure data and improve the use of third-party data by NRAs.

Since the C-Roads Platform has started, several ITS programmes have been rolled out and it has been identified that there are key elements that the NRAs will need to understand before implementing these systems more widely. The TIARA project has been designed to address the two key areas of Trust and Privacy in C-ITS applications. The first subject Trust concerns an understanding of the implementation of trust models that could protect C-ITS data. The second subject Privacy concerns an understanding of the impact of processed user personal data, including location.

Three broad research areas that have been identified:

- Trust for C-ITS applications to develop practical guidance for the implementation of PKI infrastructure for C-Roads,
- Legal and ethical ramifications for NRAs when making use of C-ITS data, and of how these change the role of the NRAs,
- Privacy impact of the processed road user location data, and recommendations to improve the location privacy-preservation for NRAs.

An experienced team of European research organisation have gathered under the coordination of AESIN/Techworkshub, the UK-based member trade association. To address this complex topic, we recognise that the best approach will be through network engagement with many organisations and individuals with experience and technical expertise, preferably independent of any specific solution vendors.

AESIN/Techworkshub belongs to the Techworkshub organisation, through which it has access to member experts in both transport and Internet-of-Things (IoT) security sectors.

SINTEF, as an independent and non-profit research organisation, has independent technical expertise and deep experience from PKI deployments in multiple sectors.

Traficon has longstanding experience of independent work with NRAs, specifically legal and ethical expertise of particular relevance to this project.

TML, bridging the gap between university and private sector, is an independent open and transparent organisation with extensive experience of data analyses and privacy ramifications.

Linking the three broad research areas identified to expertise of these organisations provides a natural project delivery structure, which will benefit CEDR and all the stakeholders involved.

A key TIARA objective is to deliver the project in close liaison with CEDR and its members, as well as the two research projects funded in the CEDR 2022 Research call on Data, Topics A (DROIDS) and B (PRESORT). The liaison ensures that results are fully compliant with CEDR and Programme Executive Board (PEB) expectations. The liaison also guarantees that the DROIDS and PRESORT projects have the possibility to utilise TIARA results and vice versa.

2.2 Purpose of this document

The main role of this work package (WP4) revolves around ensuring that road users can trust that the road operators are using data in a lawful and deliberate manner. As more data about road operators' customers is being collected by the availability of C-ITS, it becomes essential that road authorities strive to be open and transparent about it. This is especially the case for the subsequent (re)use of such data sets. Hence, road authorities should understand the types of information, characteristics, behavioural patterns, etc. that can be inferred about the customers from the data that is collected on them. To this end, this WP will help NRAs to understand the privacy impacts of the processed road user location data, as well as providing recommendations to improve the location privacy-preservation. Consequently, WP4 will address an academic review and analysis of the privacy implications of C-ITS, mainly centred around the following questions/aspects raised in the DoRN:

- Overview of research into connected vehicle re-identification or deanonymisation and research into associated preventive measures.
- Analysis of the information about road users that could be leaked from C-ITS data and the potential impact on the data subject.
- Analysis of the current measures and recommendations for additional measures to make re-identification reasonably impossible.
- Analysis of pitfalls that would increase the risk of re-identification: what to avoid in future use cases? What data cannot be added?
- Recommendations for how the understanding of the privacy provided by the system can be maintained as new use cases and as the use of data becomes more widespread.

2.3 Structure of this document

In this document we first perform desktop research of the existing literature on connected vehicle deanonymisation in Section 0. To that end, we will first provide the background for the study, after which we will provide some examples of currently encountered re-identification (i.e. deanonymisation) methods in literature, as well as possible initial countermeasures against them, closing of with an overview of several (emerging) trends in re-identification. Then, we turn our attention towards the state of the art on currently encountered mitigation measures against deanonymisation techniques. This is in turn followed by an non-exhaustive overview of the state of practice of such mitigation measures, by first giving some background against which the current mitigation measures are considered, followed by several examples of currently used and usable techniques in different domains and scenarios that are applicable for connected vehicles.

In Chapter 4 we then focus on the impact study of C-ITS data. We begin by examining the types of information that can be leaked from V2X messages, including vehicle location, speed, and heading, which pose privacy risks despite measures like pseudonymisation. We then highlight the potential for re-identification and behavioural profiling through correlations with personal data. We discuss the significant impacts on data subjects, such as the ability to infer social connections, economic status, health conditions, and enhanced personal threats from detailed movement profiles. To conclude, we underscore the necessity of evolving data protection measures, anonymisation techniques, and policies to mitigate these privacy risks, advocating for a balanced approach that combines technological innovation with robust privacy safeguards.

tbd Section 5

tbd Section 6

tbd Section 7

In addition, Section 8 contains the references to all source materials used in this report.

Finally, Appendix A in turn provides brief recapitulations of the workshops and interviews held with various experts.

The following table lists the different Expected Outcomes and the sections that address these.

Table 1: Link between Expected Outcomes and report sections.

Expected Outcome	Addressed in Section(s)
EO15 Overview of research into Connected Vehicle re-identification or deanonymisation and research into associated preventive measures.	Section 3
EO16 Analysis of the information about road users that could be leaked from C-ITS	Section 4

data and the potential impact on the data subject.	
EO17 Analysis of the current measures and recommendations for additional measures to make re-identification reasonably impossible.	tbd
EO18 Analysis of pitfalls that would increase the risk of re-identification: what to avoid in future use cases? What data cannot be added?	tbd
EO19 Recommendations for how the understanding of the privacy provided by the system can be maintained as new use cases and as the use of data becomes more widespread.	tbd

2.4 Acronyms

AI	Artificial Intelligence
ANPR	Automatic Number Plate Recognition
AT	Authorisation Ticket
BSM	Basic Safety Message
C2C-CC	Car2Car Communications Consortium
CACC	Cooperative Adaptive Cruise Control
CAM	Cooperative Awareness Message
CEDR	Conference of European Directors of Roads
C-ITS	Cooperative Intelligent Transport Systems
CPM	Collective Perception Message
DENM	Decentralised Environmental Notification Message
DGA	Data Governance Act
DoRN	Description of Research Needs
DPA	Data Protection Act
DPIA	Data Protection Impact Assessment
E2EE	End-to-End Encryption
EAA	European AI Act
EC	Enrolment Certificate
	European Commission
EDPB	European Data Protection Board
ETSI	European Telecommunications Standards Institute
EU	European Union
FOT	Field Operational Test
GDPR	General Data Protection Regulation
GNSS	Global Navigation Satellite System
IEEE	Institute of Electrical and Electronics Engineers

IoT	Internet-of-Things
ISO	International Standards Organisation
ITS	Intelligent Transport Systems
IVIM	In-Vehicle Information Message
MAP	Map Message
MAPEM	Map Data Extended Message
MCDM	Multimedia Content Dissemination Message
MCM	Manoeuvre Coordination Message
MFA	Multi-Factor Authentication
NRA	National Road Authority
OEM	Original Equipment Manufacturer
PEB	Programme Executive Board
PECR	Privacy and Electronic Communications Regulations
PKI	Public Key Infrastructure
PSM	Personal Safety Message
Re-id	Re-identification
RTTI	Real-Time Traffic Information
SAE	Society of Automotive Engineers
(S)MPC	(Secure) Multi-Party Computation
SPAT	Signal Phase and Timing Message
SPATEM	Signal Phase and Timing Extended Message
SREM	Signal Request Extended Message
SRTI	Safety-Related Traffic Information
SSEM	Signal request Status Extended Message
TIARA	Trusted Integrity and Authenticity for Road Applications
TLS	Transport Layer Security
V2X	Vehicle-to-Anything
VIN	Vehicle Identification Number

VLBS	Vehicular Location-Based Services
WP	Work package
ZKP	Zero-Knowledge Proof

2.5 Terminology and nomenclature

In light of the different concepts that are related to data privacy protection, we hereby provide some explanatory terminology and nomenclature:

- **Anonymisation** is the process of removing or modifying personal information from a dataset so that individuals cannot be identified directly or indirectly. Once data is anonymised, it should be impossible to trace back to the original individual, making the data no longer subject to data protection laws (cf. GDPR).

- **De-anonymisation** is the process of reversing anonymisation. It involves re-identifying individuals from anonymised data by using additional information or advanced techniques. This process poses significant privacy risks, as it can potentially reveal sensitive information about individuals.

- **Pseudonymisation** is a data de-identification technique where personal identifiers are replaced with pseudonyms or artificial identifiers. Unlike anonymisation, pseudonymised data can be re-identified with the use of additional information kept separately.

- **Identification** refers to the process of recognising an individual within a dataset. It involves matching data points to a specific person, allowing the data to be attributed to that individual.

- **De-identification** is similar to anonymisation but generally refers to techniques that remove or obscure personal identifiers. Unlike anonymisation, de-identification does not always ensure that re-identification is impossible. De-identified data may still pose some risk of being re-identified under certain circumstances.

- **Re-identification** is the process of matching anonymised or de-identified data back to the individual it pertains to.

Differences between anonymisation and de-identification
Anonymisation is a stronger form of data alteration, aiming to remove all identifiable information so that re-identification is not possible. De-identification, while similar, often leaves open the possibility of re-identification if additional information is available. Anonymised data is usually exempt from data protection laws, while de-identified data may still be subject to them depending on the risk of re-identification.

Difference between identification and re-identification
Identification involves recognising an individual in a dataset for the first time, while re-identification involves matching de-identified or anonymised data back to the individual after the fact.

- **Liability** in data privacy refers to the legal responsibilities and obligations that organisations have concerning the protection of personal data.

- **Repudiation** in the context of data privacy is the ability to deny the authenticity of data. It involves ensuring that a user cannot deny having performed a particular action.

- **Linkability** refers to the potential for connecting multiple pieces of data to an individual or across different datasets. High linkability increases the risk of identifying individuals, even from anonymised or de-identified datasets.

- **Unlinkability** is the property that ensures that different pieces of data cannot be linked to the same individual. It prevents the correlation of data points that could lead to identification.

- **Observability** refers to the ability to monitor, track, or observe an individual's actions or data. High observability means that actions or data can be easily monitored, which can pose privacy risks.

- **Unobservability** is the property that ensures an individual's actions or data cannot be observed or monitored.

- **Trustworthiness** in data privacy involves the reliability and integrity of data handling processes. It encompasses the assurance that data is being processed, stored, and transmitted in ways that maintain privacy and security.

- **Data minimisation** is the practice of collecting only the data that is necessary for a specific purpose and retaining it only as long as necessary.

- **Privacy by design** is an approach where privacy and data protection are embedded into the development and operation of systems and processes from the outset, rather than as an afterthought. It emphasises proactive measures to ensure privacy.

- **Consent management** involves obtaining, recording, and managing users' consent for data processing activities. It ensures that individuals are informed about how their data will be used and that they have control over their personal information.

- **Data subject rights** refer to the rights of individuals under data protection laws, such as the right to access, rectify, delete, and restrict the processing of their personal data. Ensuring these rights is essential for compliance and trust.

- **Compliance** refers to the adherence to legal and regulatory requirements designed to protect personal information. Ensuring compliance involves implementing policies, procedures, and controls to manage data responsibly, safeguarding individuals' privacy rights, and avoiding legal penalties.

- **Data breach** is an incident where personal data is accessed, disclosed, altered, or destroyed without authorisation.

- **Data protection impact assessment (DPIA)** is a process to systematically analyse, identify, and minimise the data protection risks of a project or plan. It is required under the GDPR for processing activities that are likely to result in high risks to individuals' rights and freedoms.

3 Connected vehicle deanonymisation research review

In the following sections, we will first provide the background for the study, after which we will provide some examples of currently encountered re-identification (i.e. deanonymisation) methods in literature, as well as possible initial countermeasures against them, closing of with an overview of several (emerging) trends in re-identification.

This in turn then sets the stage for the next part, where we will delve deeper in to currently encountered mitigation measures against deanonymisation techniques for connected vehicles.

3.1 Background

This section provides an overview of the research that is related to re-identification (deanonymisation) of connected vehicles and associated preventive measures. As implied by C-ITS, messages are being transmitted and received between vehicles themselves, and between vehicles and infrastructure (e.g., road-side units). These message exchanges can be cellular and/or direct (cf. DSRC) in nature. As explained by the Data Protection and Privacy Working Group of C-ITS, these messages are considered as personal data. The logic behind this statement is that the information stems from the fact that messages typically contain authorisation certificates univocally associated with the sender (which closely ties in with PKI Guidance Development work done in WP2), and furthermore even more detailed location data such as headings, timestamps, and possibly also the dimensions of the vehicle under consideration.

Furthermore, from past work done by the consortium on utilising ViaPass[1] data stemming from trucks' on-board units (used in the context of road charging in Belgium), we learned that it is possible to relatively accurately identify a truck's details based on its whereabouts and travel patterns (even with the daily rudimentary deanonymisation in place). Especially the latter providing integral clues as to what type of activities the truck is carrying out. Based on such data, it is possible to define the places where trucks stop, and even park, whereby the latter provides information as to its point of origin/destination in case of repeated behavioural (travel) patterns.

In order to counter these threats on a vehicle's/user's privacy, we will investigate two different approaches:

> (1) What is the current state of the art of mitigation measures that can establish better privacy guarantees?
>
> (2) What is the current state of practice thereby highlighting the measures that are currently being taken?

One could argue that the measures grouped under (1) are currently more theoretical in nature and should – in time – lead to (2) practical implementations in the field.

[1] https://www.viapass.be/

The research identified here is mainly retrieved from publicly available (literature) sources, as well as specific actions taken and ideas generated by various stakeholders in the ecosystem, for which we will have points of interaction during the expert workshops and conducted interviews (see also Appendix A).

Finally, in the context of connected vehicles, we highlight the difference between privacy and security. They address different aspects of data management and protection:

- **Privacy** primarily concerns the control over, and use of, personal information collected by the vehicle, such as location, driver behaviour, and vehicle usage patterns. It ensures that this information is used in accordance with user expectations and legal standards, focusing on safeguarding personal data from unauthorised access or disclosure.
- **Security**, on the other hand, refers to the protection of the vehicle's systems and networks from malicious attacks, unauthorised access, and other cyber threats. This involves implementing measures to defend the integrity, availability, and confidentiality of both the vehicle's operational and informational technologies.

While privacy seeks to protect the user's personal data, security aims to protect the vehicle itself and its systems from external and internal threats, ensuring safe and reliable operation. In our research here, we focus mostly on the privacy aspects.

3.2 Overview of general re-identification methods

This section of the report delves into the challenges and methodologies associated with the re-identification of anonymised data across various contexts. Section 3.2.1 explores the vulnerability of anonymised location data to de-anonymization techniques, e.g., by demonstrating that minimal location data points can uniquely identify individuals within large datasets. In section 3.2.2 we examine potential countermeasures that can enhance privacy protections, such as suppression and generalisation strategies in vehicle location data. Finally, section 3.2.3 reviews the evolving landscape of vehicle re-identification technology, focusing on its implications for privacy and traffic management.

3.2.1 Assessing the probability of re-identification

(Pyrgelis, et al., 2018) are among some of the researchers that investigated how anonymous location data can be deanonymised using large-scale mobility traces. As mobile technology advances, various entities, including social media platforms, mobile apps, and service providers, can access user location data with varying levels of detail. Even when users attempt to anonymise their location data by using pseudonyms, the study demonstrated that it is possible to uniquely identify individuals based on a few anonymous location points, particularly when these points are less popular or shared during working hours.

Using a dataset of network events from a European mobile operator, the research demonstrates that as few as three to four anonymous location points within a day are sufficient to uniquely identify an individual's mobility trace among tens of millions of users. This identification is influenced by factors such as the popularity of the location and the time of the day the location data is shared.

- The core concept of their methodology is to determine whether a small set of anonymised location points can be uniquely matched to the detailed mobility traces in their dataset. The assumption is that if these points can be matched to a unique trace, the trace's owner can be deanonymised.
- Each anonymised location includes a spatial and a temporal component (i.e. where and when). Based on a selective sampling of the data points, they try to find matches with full mobility traces.
- Subsequently, they calculate the probability that a limited set of specific data points will uniquely match a full trace, in order to estimate the rate of successful deanonymisation.

The study also explores the effectiveness of spatio-temporal obfuscation, reducing the granularity of time[2] and location[3] data to protect privacy. Both spatial and temporal obfuscation reduced the ability to uniquely identify mobility traces, but to varying degrees depending on the level of granularity. Spatial obfuscation generally had a more significant impact on reducing identifiability compared to temporal obfuscation. For example, increasing the spatial granularity reduced the probability of matching more effectively than increasing the temporal granularity.

[2] Modifying the time intervals at which location data is recorded or reported, e.g., changing the data from being reported every 5 minutes to every 15, 30, or 60 minutes.
[3] Changing the precision of the geographic data, such as adjusting the data from exact coordinates to generalised locations that might represent broader areas like a city block or neighbourhood, e.g., by varying the spatial resolution across different scales like 0.2 km, 1 km, 5 km, 25 km, and 125 km.

The most effective obfuscation occurred when both spatial and temporal resolutions were made coarser, suggesting that a multi-faceted approach to obfuscation could be necessary to significantly enhance privacy. However, even though findings suggest that the latter spatio-temporal obfuscation can reduce the risk of identification, it is not completely effective on its own without additional privacy-preserving measures. The reasons for this are many:

- One of the fundamental insights of the research is that human mobility traces are highly unique. The study found that even a few anonymised location points can be sufficient to uniquely identify an individual's trace within a dataset containing tens of millions of users, which is also in line with work done before by (de Montjoye et al., 2013).
- While increasing the spatial granularity reduces the probability of a unique match, the study showed that traces can still be relatively unique even at higher levels of spatial obfuscation. For example, even when location data was generalised to larger areas (up to 25 km or more), a significant percentage of traces could still be matched uniquely.
- Similarly, increasing the temporal granularity also did not fully anonymise the data. The probability of uniquely identifying a user's trace decreased as the interval between data points increased, but the effect was not enough to eliminate the risk of deanonymisation entirely. Combining this with increasing the spatial granularity did not reduce the identification risk to negligible levels. This suggests that the inherent uniqueness of mobility patterns makes them difficult to disguise completely through obfuscation alone.
- Furthermore, powerful enough statistical techniques could potentially unmask individuals even when data is obfuscated. The unique patterns of movement and regularity in timing (e.g., daily commutes or frequent visits to certain locations) can still provide enough information for algorithms to re-identify individuals, especially if additional contextual information is available.
- The latter point is closely tied to the possibilities when also cross-referencing data from different sources. This poses another challenge to the effectiveness of spatio-temporal obfuscation. If an attacker can access multiple data sources that include some amount of location information, they might integrate these to circumvent the obfuscation in one or all sources.

Similarly, (Tan, et al., 2017) investigated the vulnerability of privacy in vehicular location-based services (VLBS) despite common anonymisation practices like dummy data and *k*-anonymity.

- **Dummy data** involves adding fake entries into the dataset. These dummy entries are designed to look realistic but do not correspond to any real individual. The purpose is to confuse potential data attackers or anyone trying to re-identify individuals from the dataset by increasing the difficulty of distinguishing real data from fake data. This method can be particularly useful in environments where data points (like vehicle locations) are monitored and collected, as it creates noise and reduces the accuracy of attempts to match the data with real-world identities.
- ***k*-anonymity** is a more formal privacy protection model that ensures that individuals are indistinguishable from at least $k - 1$ other individuals in the dataset. Under *k*-anonymity, the information for each person contained in the release cannot be distinguished from at least $k - 1$ individuals whose information also appears in the data set. This is typically achieved by suppressing or generalising certain identifiers (like exact locations or times in the context of VLBS) so that each set of data containing the same values in the identifying fields contains at least *k* entities. For example, rather than recording the exact location of a vehicle, the data might be generalised to a broader region that includes at least *k* different vehicles, thereby obscuring individual movements.

They found that vehicular trajectories are highly unique; notably, only four spatio-temporal points are needed to re-identify a vehicle with over 95 % accuracy, underscoring significant re-identification risks in VLBS.

Their study analysed two large datasets of taxi trajectory metadata from Shenzhen and Shanghai, encompassing over 1.1 billion records. This detailed analysis revealed that vehicles' location data, even though anonymised, can still be uniquely identified due to the constrained nature of their movements along roads, which enhances traceability and diminishes the effectiveness of anonymisation techniques.

The authors emphasised that even with anonymised datasets, the uniqueness of vehicular movements makes privacy preservation challenging. They suggested that more robust anonymisation methods are necessary to truly safeguard user privacy in VLBS. To this end, they proposed several strategies for both users and service providers to enhance privacy protection, including minimising the use of location services and employing more sophisticated cryptographic methods.

(Gao, et al. 2019) explored privacy concerns associated with publishing individual-based mobility trace data, particularly from automatic number plate recognition (ANPR) systems. Similar to (Tan, et al., 2017), they employed the concept of *k*-anonymity to measure the risk of privacy breaches via re-identification attacks. Analysing a month-long ANPR dataset from Guangzhou (comprising 260 million spatiotemporal transactions from 14 million vehicles), the researchers identified factors that influence anonymity, such as data size, temporal granularity, and local versus non-local vehicles.

Their findings showed that having as few as five spatiotemporal records was enough to uniquely identify about 90 % of the individuals in the dataset, even when the temporal granularity was set to half a day. This illustrates how only a few data points are necessary for a potential breach of privacy. In similar vein, they explored how the temporal resolution of data impacts privacy. They demonstrated that increasing the granularity of the temporal data (i.e. from minutes to hours) increases anonymity but does not eliminate the risk, as a significant portion of the dataset could still be uniquely identified. Further analysis revealed differences in anonymity between local and non-local vehicles, indicating that even the geographic context of data points affects re-identification risks. Local vehicles tended to have higher anonymity levels, likely due to more frequent and diverse movements within the dataset's coverage area, contrasting with non-local vehicles that might follow predictable patterns when entering or leaving the area.

As such, even minimal spatiotemporal records can uniquely identify a large percentage of individuals, underscoring significant privacy risks.

3.2.2 Possible initial countermeasures

(Carter and Ferber, 2019) introduced a novel de-identification procedure for location data from connected vehicles. The primary concern addressed is the vulnerability of this data to inference-based privacy attacks which could potentially lead to the identification of individuals based on vehicle location patterns. More traditional anonymisation techniques like location generalisation or perturbation are deemed inappropriate for connected vehicle data, as they can degrade the data's utility for safety-critical applications that require precise position information.

The authors propose a suppression-based method that utilises the structure of the road network to protect location data. Their approach aims to balance privacy protection and data utility without significantly reducing the data's effectiveness for applications like vehicle-to-vehicle (V2V) communication. The method was tested using data from the first U.S. deployment of a connected vehicle model, which involved over 460,000 vehicle trips and nearly 4 billion GPS points.

- Firstly they accurately map match each data point to a known road segment (using Open Street Map), which is then followed by understanding the intersections and the paths taken by vehicles in order to identify how many potential routes a vehicle could have taken to reach a point. Here, ambiguity is added to the exact path travelled by any vehicle.
- Critical intervals (areas near sensitive locations like homes) are identified, and privacy intervals are then established to obscure these critical points by suppressing nearby location data.
    - **Critical intervals**: these are specific segments of a trip that are deemed sensitive due to their proximity to locations like a person's home or workplace. These are the areas where a vehicle might be stationary for significant periods (like homes, schools, or workplaces), or they are typical start or end points of a journey. Identifying these intervals is crucial because they are most likely to contain privacy-sensitive information that could lead to the reidentification of an individual if exposed.
    - **Privacy intervals**: once critical intervals are identified, the algorithm establishes extended segments of the trip data surrounding and including the critical intervals. The purpose of privacy intervals is to obscure the exact locations within the critical intervals by also suppressing data points around them. The extension of suppression beyond just the critical points helps prevent attackers from deducing sensitive locations by examining only the surrounding data points.

Suppressing the data as such in privacy intervals then involves dynamically adjusting their size based on road network characteristics like the number of potential exits and entries at intersections within the suppressed segment. This increases the entropy, or unpredictability, of the data set, which makes it harder for an adversary to confidently infer the true paths or locations of individuals. Importantly, the parameters for defining the size of privacy intervals are chosen to strike a balance between obscuring sensitive locations and retaining enough data utility for analysis and safety applications.

By effectively using the road network structure to identify and suppress critical and privacy intervals, the method enhances privacy protection without significantly diminishing the data's value for applications that depend on accurate and comprehensive location data. This approach allows for the continued use of connected vehicle data in developing safety measures and traffic management solutions while safeguarding individual privacy.

(Tan, et al., 2017) provided several strategies and recommendations to enhance privacy protection for both users and service providers of vehicular location-based services. Some of these are more applicable or feasible then others.

- **For users**
  - Users should use VLBS-applications only when necessary, as these services continuously upload location data, which could compromise privacy.
  - Users should avoid providing unnecessary personal information when registering or using LBS-related applications, reducing the risk of personal data being associated with their location data.
  - Using different accounts for different services or even different applications for similar services can help in dispersing location traces and confusing potential trackers.
- **For VLBS providers**
  - Providers should consider reducing the frequency of information collection if it does not affect the quality of service, thereby reducing the amount of location data stored and potentially exposed.
  - Employing stronger cryptographic methods can help in securing stored data, making it more difficult for unauthorised parties to access or decipher it.
  - Storing minimal amounts of sensitive information can reduce the impact of a data breach and help maintain user trust.

These strategies aim to address the high re-identification risk highlighted by the study, suggesting that while conventional anonymisation techniques like *k*-anonymity are in use, the unique nature of vehicular movements requires more tailored and robust approaches to ensure privacy in VLBS.

(Gao, et al. 2019) proposed two methods to enhance data privacy while maintaining utility:

- **A suppression solution that identifies and removes sensitive records**
  A sensitive record was defined as any record with the number of vehicles that shared the same spatiotemporal point (or closely similar points) less than a predefined threshold. The sensitivity of a record was also assessed based on traffic volume data, whereby records corresponding to lower traffic volumes were more likely to be classified as sensitive because they inherently included fewer vehicles and thus smaller anonymity sets.
  Once sensitive records were identified, they were removed from the dataset. The researchers then evaluated the trade-off between increased privacy (higher anonymity) and the loss of data utility. They quantified the data loss by measuring how much of the dataset was removed and assessed whether the removal significantly impacted the utility of the remaining data. The researchers explored whether the reduction in dataset size compromised its usefulness for analysis.
  The effectiveness of the suppression solution was then assessed by measuring the increase in anonymity across the dataset after the sensitive records were removed. Similar to the approach by (Carter and Ferber, 2019), they examined the balance between enhancing privacy and retaining data utility. The suppression solution effectively increased the anonymity of the dataset by removing records most vulnerable to re-identification attacks. For example, with certain parameter settings, the average individual anonymity identified by three spatiotemporal records increased by more than 20 %. Notably, this increase in anonymity was achieved with less than an 8 % loss of data, indicating a favorable balance between privacy protection and data utility.
- **A bintree-based generalisation solution**
  This solution started with a high-resolution temporal data set where each record's timestamp was quite precise. Then, to generalise this data, the solution progressively merged temporal intervals to increase their size until each newly formed interval met a predefined minimum anonymity threshold (similar to the *k*-anonimity logic).
  Next, the solution used a binary tree structure, which provided flexibility in how time intervals could be split or merged. This structure allowed the algorithm to adaptively adjust the boundaries of time intervals based on traffic volume and anonymity requirements. The key goal was to achieve the desired level of anonymity with the least possible loss of information. The algorithm determined optimal points to split the time intervals by minimising an empirical measure of information loss, essentially preserving as much original data utility as possible. Statistically speaking, information loss was quantified using an entropy-based measure, reflecting the loss of data granularity and detail due to the generalisation of time intervals. The solution furthermore dynamically adjusted the length of time intervals based on real-time traffic data to ensure that each interval maintained the minimum required anonymity while striving to minimise the loss of useful information.
  Their results showed that the bintree approach, compared to a standard time interval cloaking approach, was more effective in balancing anonymity with data utility (by measuring an increased average individual anonymity with minimal loss of data), particularly in how it adapted to varying traffic conditions throughout the day, and that the method scaled efficiently with large datasets. This bintree-based generalisation technique has parallels in fields like social network analysis and epidemiology, where

balancing privacy and data utility is critical. In the former, similar methods have been applied to anonymise user interactions while preserving key relationships, often using *k*-anonymity and entropy-based measures to minimise information loss (Zheleva and Getoor, 2007). Likewise, in the latter, protecting patient privacy in time-sensitive health data is crucial, and adaptive interval generalisation helps maintain confidentiality without sacrificing important insights about disease progression (Malin and Sweeney, 2004). These applications validate the effectiveness of bintree-based approaches in preserving privacy while ensuring data utility across diverse domains.

On the other side of the spectrum, we can look at techniques that can be used to perform vehicle re-identification, and then extract potential countermeasures against this. To start from, (Zakria, et al., 2021) provided a comprehensive overview of trends in vehicle re-identification in camera-based systems. The authors introduced vehicle re-identification as a crucial component of intelligent transportation systems. In their approach, this involved recognising a vehicle across different cameras within a non-overlapping surveillance network, posing challenges due to variability in vehicle appearance, camera viewpoints, and environmental conditions.

Here, several challenges hinder the effective implementation of vehicle re-identification systems, such as:

- **Inter-class similarity and intra-class variability**: different vehicles may look similar, and the same vehicle may look different under various conditions.
- **Viewpoint changes**: Vehicles captured from different angles present significant recognition challenges.
- **Spatio-temporal uncertainty**: The time and location of a vehicle capture can affect identification accuracy.

Based on this, their paper discussed various approaches to vehicle re-identification, including (i) vision-based methods (focusing on analysing visual features such as vehicle model, colour, and number plates), (ii) sensor-based methods (using magnetic, inductive loop sensors, or GPS to track vehicles based on physical or positional characteristics), and (iii) hybrid approaches (combining multiple data sources and technologies to improve identification accuracy).

Moreover, the review provides a comparative analysis of the performance of current state-of-the-art methods using the VeRi-776 and VehicleID datasets, aiming to provide future research directions in vehicle re-identification. This included a discussion on the significance of vehicle re-identification in managing traffic congestion, reducing carbon dioxide emissions, enhancing road safety, and supporting the overall complexity of transportation systems.

In order to defend against these vehicle re-identification solutions, a number of options are available, with a strongly varying degree of applicability and feasibility. As such, countermeasures (i.e. mitigation measures) to safeguard privacy and security can include the following:

- **Vehicle appearance modification**: regular changes to the physical aspects of vehicles can help evade re-identification. This could include temporary or interchangeable vehicle wraps that change the colour and patterns of the vehicle exterior, and modular physical components such as different mirrors, spoilers, or hubcaps that can be swapped regularly. Additionally, adversarial patches, i.e. small, strategically placed stickers or designs, can be used to confuse automated recognition systems, making it harder for them to correctly identify or track vehicles.
- **Number plate masking technologies**: these include technologies that allow for the digital or physical masking of number plates in certain scenarios, e.g., electronic number plates that can change numbers or become unreadable outside of legal monitoring areas, or physical covers that are legal and can mask the plate when the vehicle is parked or in specific private areas. Some of these techniques are more oriented towards counter surveillance, such as the use of reflective paints and materials that interfere with camera detection and recognition capabilities, or infrared led systems around the number plate or the vehicle that obscure the vehicle's features from cameras but are invisible to the human eye. In similar spirit, classic GPS jammers can act as devices that disrupt or block GPS signals, preventing accurate tracking of the vehicle's location. Note however that these kinds of solutions are challenging to implement within a given legal context.
- **Controlled disclosure of location data**: here the idea is to implement systems that control when and how vehicle location data is shared, such as vehicle tracking opt-out systems for users who do not wish their data to be shared, and encrypted GPS data that can only be accessed by authorised entities under specific conditions.
- **Legal and regulatory measures**: the goal of these is to strengthen the legal framework to protect individuals' privacy by adopting laws that regulate the use and scope of vehicle re-identification technologies, as well as data-minimisation principles that require only essential data collection for specific, lawful purposes.
- **Decentralised vehicle identification systems**: these are aimed at reducing centralised tracking by utilising, e.g., block chain technology for vehicle registration and identification that ensures data integrity and security while allowing for anonymity, and peer-to-peer network systems for vehicle communication that do not require a central monitoring authority.
- **Standardisation of anonymisation protocols**: it is important to create and adopt standards for anonymising collected data (e.g., with hashing and/or encryption) to ensure that data cannot be reverse-engineered to reveal identities, and secure multi-party computation techniques[4] that allow data processing without exposing the underlying data.

[4] In short, secure multi-party computation is a cryptographic technique that enables multiple parties to jointly compute a function over their inputs while keeping those inputs private. The parties are able to derive a correct output without revealing their individual data to each other or to any other entity. This

Implementing these countermeasures requires a balance between enabling the technological benefits of vehicle re-identification for safety and efficiency, and protecting individual privacy rights. **Effective policies and technological solutions need to be developed collaboratively by governments, industry stakeholders, and privacy advocates** to ensure that vehicle re-identification technologies are used ethically and responsibly.

method ensures that each participant's data remains confidential, and only the specific outputs or computations agreed upon are shared.

3.2.3 General trends in re-identification

The scientific literature on vehicle re-identification has evolved over the years with distinct trends reflecting advances in technology, computational methods, and the growing needs of applications such as traffic management, security, and automated driving. In this section we provide some of the key trends that characterise the field.

Over the past decade, there has been a significant shift towards using deep learning techniques for vehicle re-identification. These methods have shown superior performance over traditional machine learning approaches due to their ability to learn complex, hierarchical features from large amounts of data. Convolutional neural networks and other deep architectures are commonly employed to handle variations in lighting, angle, occlusion, and vehicle modifications. An example of such more sophisticated techniques is the work of (Lian et al., 2022) with transformer-based attention networks. These networks utilise spatial attention mechanisms[5] to learn discriminative features essential for vehicle re-identification. They work by focusing on specific parts of a vehicle that are invariant to changes in viewpoint or illumination, such as colour or car type. This method helps in accurately identifying vehicles from different angles and lighting conditions, thereby improving the re-identification process while potentially enhancing privacy by focusing on less distinctive features.

In addition, there is a growing trend to enhance re-identification accuracy by integrating multiple data modalities. This includes combining visual data with other forms of data such as LIDAR, radar, or metadata (e.g., time stamps, GPS coordinates). The integration of these data types helps mitigate the limitations posed by visual-only data, especially in challenging environmental conditions. Advanced methods now incorporate spatial and temporal information to improve re-identification accuracy. Techniques that track the movement of vehicles over time and across different camera views help establish stronger identity recognition, especially in dense traffic scenarios and over large surveillance areas. An example of this is the PROVID framework by (Liu et al., 2018), which – in the context of number plate recognition systems – begins with a coarse search that narrows down potential matches before refining the search based on more detailed attributes. Their method leverages multimodal data, including visual features and contextual information like number plates and camera locations, to improve accuracy and efficiency.

Note that as the availability of annotated datasets varies widely across different regions and conditions, there is an emphasis on cross-domain adaptation techniques that allow models trained on one dataset to perform well on another, without extensive retraining. Related to this we see a rise of recent studies that employ attention mechanisms and feature fusion strategies to focus on the most relevant features of a vehicle for identification purposes. This can include specific parts of a vehicle like the number plate, make/model logos, or distinctive design features, which are crucial for distinguishing between similarly-looking vehicles.

[5] Attention mechanisms in machine learning dynamically prioritize different parts of input data to enhance model performance, particularly in tasks involving sequences or spatial data.

In a broader context, we observe, somewhat counterintuitive to the general trend of this deliverable, that with the rise of smart cities and real-time traffic management systems, there is a pressing need for real-time vehicle re-identification. This arises from the increasing demand for advanced traffic management, security, and automation in smart cities. As cities integrate real-time systems to monitor and control traffic flows, vehicle re-identification becomes critical for tracking vehicles across different locations, ensuring smooth traffic flow, and responding to incidents swiftly. Research is increasingly focused on developing algorithms that not only provide high accuracy but also meet the latency requirements for real-time applications.

Furthermore, as the deployment scenarios become more complex, the robustness and scalability of vehicle re-identification systems are under continuous improvement. This includes enhancing the system's ability to handle large-scale deployments across multiple locations and maintaining high performance in diverse weather and lighting conditions.

Of course, as vehicle re-identification technology is closely linked to surveillance, there is an ongoing discussion regarding privacy and ethical implications. The literature often addresses the need for balancing technological advancements with privacy rights, advocating for regulations and frameworks that protect individual privacy while allowing for the beneficial uses of re-identification technologies. This striking of a balance between protecting a user's privacy versus the ability to still provide tracking and extraction of useful information (for policy-based purposes) remains always at the forefront of any discussion.

3.3 State of the art of mitigation measures

Addressing the concerns surrounding privacy in the context of connected vehicles is critical as these technologies gather and transmit vast amounts of data that can potentially reveal sensitive personal information. The state of the art for privacy under the umbrella of connected vehicle deanonymisation involves a combination of technical measures, legal frameworks, and organisational practices. In turn, we will discuss the following different types of measures:

- Anonymisation, access control, and data minimisation
- Differential privacy and synthetic data
- (Homomorphic) encryption
- Secure multi-party computation
- Zero-knowledge proofs
- Federated learning
- Legal frameworks and regulations

3.3.1 Anonymisation, access control, and data minimisation

Reducing the amount of personal data collected to the minimum necessary can help mitigate risks. Anonymisation and pseudonymisation techniques alter data so that individuals cannot be identified without additional information. Anonymisation is essential to protect privacy when sharing data, yet it is difficult to achieve if the original dataset remains accessible. We propose that legal definitions require that anonymised data cannot be traced back to individuals using common methods. The GDPR defines pseudonymisation as removing direct identifiers and replacing them with pseudonyms (using hashing or creating or tokens). This is a common practice, where data is adjusted to minimise the risk of re-identification. The levels of de-identification vary, with more advanced levels offering greater privacy protection. The challenge with pseudonymisation however is its limited ability to ensure complete privacy, because it does not modify other indirect data that could lead to re-identification. The effectiveness depends on how isolated the pseudonymised data remains from the original identifiable dataset, which can be difficult to manage, especially in complex data environments where multiple parties share data. Balancing the utility of data with privacy protection remains a significant challenge, as pseudonymisation does not guarantee anonymity if the original dataset is accessible, as stated before.

Using decentralised identity solutions where identities are not centrally stored can reduce the risk of mass data breaches. Furthermore, robust anonymisation of data prior to transmission or storage ensures that even if data is intercepted, it cannot be linked back to an individual without significant additional information. In light of this, some block chain applications can enhance privacy by decentralising data control and enabling transparent, secure, and tamper-proof systems. In a nutshell, block chain technology offers significant potential for enhancing privacy protection through its inherent characteristics of decentralisation, transparency, and security. Specifically applicable to connected vehicles, these systems can utilise decentralised identity solutions where vehicle and user identities are managed without a central repository, significantly reducing the risk of mass data breaches as there is no single point of failure. Additionally, block chain's capacity for maintaining a tamper-proof ledger ensures that data transactions (such as location sharing, usage statistics, and vehicle status updates) are

recorded securely and immutably. This not only aids in robust anonymisation techniques – by verifying and enforcing privacy rules before data is transmitted or stored – but also enables vehicle owners to have transparent visibility and control over who accesses their data. This approach helps to prevent unauthorised access and ensures that data, even if intercepted, cannot be linked back to individuals without considerable additional information, thereby safeguarding personal privacy.

In light of authorising access, we note, in conjunction with the work done in WP2 of this study, the use of PKI for secure communication between connected vehicles and other ITS components (ETSI, 2022). This also involves the specification of roles like an Enrolment Authority and an Authorisation Authority to manage credentials and access controls effectively. Here, privacy necessitates the need for unlinkability, ensuring vehicles can communicate securely without unnecessary identity exposure. An example of this is the Flemish Mobilidata programme (AFB, 2020). There, the importance of confidentiality, integrity, and authentication in vehicular communications is stressed by using PKI. Specific privacy measures include encryption to protect data confidentiality, mechanisms to ensure data integrity, and robust authentication to allow only authorised devices to communicate within the network.

An important aspect in this view is allowing users to manage who has access to their data and for what purpose, potentially in real time, provide a higher degree of control and privacy. Here, dynamic consent frameworks are highly relevant for enhancing privacy by providing drivers and vehicle owners with continuous and granular control over their personal data. In contrast to static consent, where permissions are given once (typically at the time of purchase or service initiation), dynamic consent allows for ongoing adjustments to consent based on the context and user preferences. This model is particularly advantageous in connected vehicles, where data types and usage can vary significantly, ranging from real-time location tracking to usage patterns and vehicle health data. By implementing a dynamic consent framework, vehicle manufacturers and service providers can empower users to decide what data they are comfortable sharing, with whom, and under what circumstances. This not only enhances trust and transparency but also aligns with stricter data privacy regulations, ensuring that personal data is handled in a manner that respects user privacy preferences at all times. Similarly, techniques such as role-based access control (e.g., assigning permissions to users based on their role within an organisation.) and attribute-based access control (evaluating attributes associated with a subject, object, requested operations, etc., in order to determine the subject's authorisation to perform a set of operations) also provide higher degrees of security.

Related to this is the use of data spaces. These are virtual architectures designed to facilitate the secure and controlled exchange of data. Framework such as these help protect privacy by establishing strict data governance rules and enabling data operations that comply with legal and security standards. For connected vehicles generating and processing vast amounts of data, including sensitive information like location, driving patterns, and personal preferences, data spaces ensure that this data can be shared with manufacturers, service providers, and third-party applications in a manner that prioritises user consent and privacy.

Finally, there exist specific methods, like video coding, that involve sharing only essential data elements (like codes and timestamps) rather than raw data (including the original video footage), which helps protect privacy while retaining useful information[6]. These data reduction techniques reduce the risk of re-identification. In addition, they also help manage the volume of data, facilitating easier handling and storage. Plus, such processes of feature extraction can preserve privacy while maintaining data utility. This is critical as datasets often need to be purged of personal data after use due to legal and ethical restrictions.

Note that, in order to instigate trust, it is necessary that all these systems are made transparent about what data is collected, how it is used, and who it is shared with. Tools that provide users with clear visibility and control over their data can empower them to make informed decisions about their privacy.

[6] This is in essence the approach that Telraam uses (see also https://telraam.helpspace-docs.io/article/29/what-about-privacy).

3.3.2 Differential privacy and synthetic data

Differential privacy as elaborated by (Dwork et al., 2006a) is a sophisticated privacy-enhancing technique that introduces controlled randomness into the data itself (or into the functions processing the data), ensuring that individual information remains obscured even when aggregated datasets are analysed or shared. By adding random noise to the data before it is shared, it becomes considerably more difficult to trace specific data points back to individual vehicles or drivers. This is particularly useful when vehicle data needs to be shared with third parties for purposes such as traffic management. The noise here refers to random data added to actual data points to obscure the values of individual entries, or to the results of queries run on the data, depending on whether the approach is local or global differential privacy. This noise is typically generated from a probability distribution, such as Laplace (i.e. double exponential) or Gaussian (i.e. normal) distributions. These distributions are commonly used in differential privacy due to their mathematical properties that align well with the privacy guarantees required. E.g., the Laplace distribution is favoured because of its sharp peak and heavy tails, which make it effective for adding the type of noise that can mask individual entries in a dataset while still allowing for accurate aggregate information to be derived. It provides a mechanism to adjust the scale of noise directly proportional to the sensitivity of the data. A Gaussian distribution is used when a small probability of privacy failure is permissible. Gaussian noise is also smoother and can be beneficial when dealing with data requiring stricter control over the tails of the distribution, which reduces the risk of extreme values that might inadvertently reveal sensitive information.

The effectiveness of differential privacy in protecting privacy while retaining utility in the data is notable, but it also comes with caveats. The key is in balancing the amount of noise added: the more noise introduced, the greater the privacy but at the potential cost of the usefulness of the data. For instance, excessively obscured data lowers the data's utility and even may lead to inaccurate traffic forecasts or inefficient urban planning. Therefore, implementing differential privacy requires careful calibration to ensure that the data remains useful for analysis without risking individual privacy. This balance is critical in maintaining the trust of vehicle users and the utility for third-party analysts.

The technical implementation of differential privacy involves algorithms that are designed to aggregate data in a way that any single data point (or user) does not significantly influence the outcome of the analysis. This means that the presence or absence of any individual in the dataset does not alter the overall data significantly, thereby preventing the possibility of identifying that individual through reverse engineering or other data analysis techniques. For connected vehicles, this might involve complex data processing systems embedded within vehicles' telematics or carried out at the data aggregation stage by third-party processors. Note that this technical challenge is substantial, as it requires not only the development of robust privacy-preserving algorithms but also their integration into the existing vehicle and data infrastructure. Moreover, maintaining and updating these systems to cope with evolving data types and increasing volumes, while ensuring compliance with global data protection regulations, adds another layer of complexity to the use of differential privacy in connected vehicles.

Regular so-called ϵ-based differential privacy, also known as pure differential privacy, is a stringent privacy standard that relies solely on the epsilon parameter to provide a strong guarantee of privacy. This parameter quantifies the degree of privacy protection by limiting the amount of information that can be inferred about any individual in the dataset from the output. The smaller its value, the higher the level of privacy, as it signifies a smaller probability that an observer can distinguish whether any individual's data was included in the dataset based on the output. This approach does not allow for any probability of failure beyond what is bounded by epsilon. However, while offering robust privacy, pure differential privacy can sometimes be too restrictive, leading to a significant compromise on the utility or accuracy of the data, especially in complex or sensitive data applications where precise data analysis is crucial. Note that, as explained by (Mehner et al., 2021), there is some complexity involved in interpreting the epsilon parameter. In their research they introduced an improved model that simplified the understanding of epsilon by focusing on worst-case scenarios, thus making the parameter more accessible to both data engineers and data subjects. They present global privacy risk and leak concepts with clear mathematical definitions, helping stakeholders evaluate privacy safeguards effectively. The authors also advocate an emphasis on the need for better communication strategies to explain the implications of the epsilon parameter.

To remediate this somewhat, the framework of (ϵ,δ)-based differential privacy, also known as approximate differential privacy, extends the traditional model of differential privacy by introducing an additional parameter delta that allows for a small probability of the privacy guarantees being breached, as explained by (Dwork et al., 2006b). This framework provides a more flexible approach to privacy that can be particularly useful when dealing with complex or high-dimensional data sets where the earlier mentioned pure differential privacy might be too limiting or impractical. The addition of the delta parameter acknowledges that with a probability no greater than delta, the differential privacy guarantee may not hold. This essentially allows the privacy mechanism to have a small probability of failing to completely anonymise the data, typically in situations where a strict guarantee would require adding an impractical amount of noise. In any case, this parameter is typically set close to zero, indicating that the probability of such a privacy breach is extremely low, but non-zero. This dual-parameter approach balances the need for practical data utility with robust privacy protections, enabling analysts to manage the trade-off between data accuracy and privacy in scenarios where perfect anonymity is challenging to achieve.

Both pure and approximate differential privacy can be applied to address various data privacy challenges while enabling the useful analysis of collected data. Consider for example the following use cases for pure differential privacy:

- **Location data aggregation**: pure differential privacy could be applied to the process of aggregating location data from multiple vehicles to create heat maps of traffic density. By adding noise to the counts of vehicles in each geographical grid cell, individual vehicles' locations remain obscured, thus protecting privacy. The strict control of the epsilon value ensures that it is not possible to determine whether a specific vehicle was in a particular location, maintaining strong privacy guarantees even while allowing the production of useful traffic flow insights, for example as done by Google in (Eland, 2015).

- **Driving behaviour analysis**: one might use pure differential privacy to analyse driving patterns across different regions without compromising the privacy of individual drivers. By applying noise to metrics like average speed, braking habits, or fuel efficiency before they are shared or analysed, researchers can still draw conclusions about general driving behaviours without risking re-identification of the data sources.

Similarly, there are relevant use cases for approximate differential privacy:

- **Machine learning models**: here the technique is particularly useful when training machine learning models on large datasets. In this case, a small probability of privacy leakage is acceptable to ensure that the models are sufficiently accurate. For example, models predicting vehicle maintenance needs or optimising route efficiency can be trained on aggregated data with a slight relaxation in privacy to retain more detailed patterns in the data that are crucial for accurate predictions.
- **Real-time data sharing**: in scenarios where connected vehicles share data in real-time with traffic management systems, e.g., to optimise traffic flows or reduce congestion, approximate differential privacy can be employed. It allows for a practical level of noise addition, i.e. less than what pure differential privacy would necessitate, thereby maintaining higher data utility for real-time decision-making processes while still offering substantial privacy protection.

These examples show how differential privacy can be tuned to the specific needs and risks associated with different types of data usages in connected vehicles, balancing the trade-offs between data utility and privacy.

In similar vein, adding data by fabricating synthetic data and inserting into the original data, may also seem a useful technique. However, as explained by (Kapp et al., 2023a) and (Kapp and Mihaljevic, 2023b), the generation of synthetic data in this context becomes a significant challenge due to the detailed and sensitive nature of the data collected, which can include vehicle locations, driver behaviour, and traffic patterns. Synthetic data generation aims to create data sets that structurally and statistically mirror real-world data but without compromising sensitive information. This process helps in addressing privacy concerns and facilitates the open sharing of data. Generating such synthetic data is complex due to the high dimensionality and sparsity of the data. These synthetic data generation algorithms, despite their theoretical advantages in privacy preservation and data sharing, still often fail to deliver realistic and practical outputs when tested against real-world conditions. The research underscores that most algorithms struggle with replicating authentic traffic patterns, vehicle interactions, and the dynamic nature of connected vehicle networks. This gap between theoretical utility and practical applicability in the current state of the art suggests a need for more refined models that can better capture the complexities of modern transportation systems.

A promising avenue that still requires further exploration is geo-obfuscation, whereby deliberate alterations or degradations to the accuracy of geographic information of vehicles' locations is done to prevent exact location tracking while still providing useful data for applications that require some level of location information. First, the actual geographic coordinates are modified by adding a random amount of noise to the location data. This alteration is controlled to ensure that the location remains useful for certain applications but is imprecise enough to protect the user's privacy. Then, the granularity of location data is increased, such as reporting the location at a block or neighbourhood level instead of precise coordinates. The location is generalised to a larger geographic area that includes the actual location, ensuring that the individual or vehicle cannot be singled out. Note that geo-obfuscation can also be implemented by defining safe zones where location data is either not collected or heavily obfuscated. This is common in areas where privacy concerns are particularly sensitive, such as around homes or personal destinations.

3.3.3 (Homomorphic) encryption

Advanced encryption plays an important role in enhancing data security, especially in contexts where sensitive information needs robust protection. End-to-end encryption is one such method where data is encrypted at the source and remains encrypted until it reaches the intended recipient, who then decrypts it using a specific key. This ensures that the data remains unreadable by any intermediaries (eliminating so-called man-in-the-middle attacks). With end-to-end encryption, data such as location, vehicle status, or diagnostic information can be securely transmitted between the vehicle and the manufacturer or service providers, ensuring that only authorised parties can access and read it.

Note that in the world of encryption, we broadly distinguish between symmetric and asymmetric encryption. The former uses the same key for encryption and decryption (examples are Data Encryption Standard, DES, and Advanced Encryption Standard, AES), while the latter uses a pair of different keys, i.e. a public and a private one, for both encryption and decryption (an example is Pretty Good Privacy, PGP). The public key fulfils the role of providing trustworthiness, while the private key provides for integrity and repudiation such as for digital signatures (De Vuyst et al., 2022).

Homomorphic encryption represents another significant advancement in encryption technology, as elaborated by (Boudguiga et al., 2021). It allows for data to be processed while still encrypted, which means that it can be analysed, manipulated, or transformed without ever exposing the actual underlying data. This type of encryption maintains the confidentiality of the data throughout its lifecycle, not just when it is stored or being transmitted. For instance, homomorphic encryption enables third parties to perform complex calculations on encrypted data without needing access to the raw data. Homomorphic encryption allows for vehicle data to be analysed by third parties, e.g., traffic management systems or insurance companies, without ever having to decrypt it. Thus, insights related to traffic patterns, driving behaviours, etc. be extracted while preserving the anonymity and privacy of the individual drivers.

These techniques enable a safer integration of connected vehicles into wider data-driven systems like smart city infrastructures, where data can be utilised for broader benefits without compromising individual privacy. Thus, encryption not only secures data against unauthorised access but also enhances the feasibility of sharing and analysing vehicular data in a privacy-preserving manner.

Note however that, as (Case, 2023) explains, depending on the type of homomorphic encryption, it may not support all computations that are needed by organisations. And even when it does, it may require significant computational overhead to perform intensive calculations on encrypted data, making it either slower or more resource intensive. This is a big challenge for homomorphic encryption and as such, the technique may be less practical for NRAs.

3.3.4 Secure multi-party computation

(Secure) multi-party computation (MPC) is a cryptographic technique that allows multiple parties to collaboratively compute a function over their inputs while keeping those inputs private, as explained by (Zhao et al. 2019). Each participant in the computation contributes their piece of data, which is combined to produce a result, such as a sum or average, without any party revealing their individual inputs to the others. This method is based on complex cryptographic protocols which ensure that, although the inputs are used in the computation, they are never exposed to other participants. MPC is especially valuable when sharing the raw data is too sensitive or when privacy needs to be strongly preserved, even while deriving joint insights.

MPC can be extremely useful for applications that require pooling data from multiple vehicles to enhance safety or traffic efficiency. For example, vehicles could use MPC to calculate the average speed of traffic in a particular area without any single vehicle having to reveal its speed to others or a central server. This approach not only helps maintain the privacy of the individual vehicle's data but also enriches the collective utility of the data shared among vehicles. Additionally, MPC can enable real-time decision-making in traffic management systems by aggregating data like location or congestion levels, all while safeguarding the privacy of each data contributor.

Beyond traffic and safety applications, MPC can also facilitate cooperative interactions between vehicles and urban infrastructure without compromising the privacy of the data involved. For instance, vehicles could communicate with traffic lights or parking management systems to optimise routing and parking solutions based on real-time data analysis performed via MPC. This means that a vehicle could contribute data to a system that calculates the optimal timing for traffic lights in an area or the best distribution of parking spaces among current users, without revealing any individual user's location or destination.

MPC enables vehicles to contribute to collective data processes without exposing individual data, which is especially crucial as the balance between utility and privacy becomes increasingly cumbersome. MPC's ability to compute over encrypted or anonymised data ensures that connected vehicles can participate in broad data-driven initiatives without risking the privacy of the vehicle owners or the integrity of their data. A technique such as MPC can therefore be more suitable and feasible as a solution for NRAs.

3.3.5 Zero-knowledge proofs

There exist methods by which one party can prove to another that a given statement is true, without conveying any additional information apart from the fact that the statement is indeed true. These so-called zero-knowledge proofs (ZKPs) are advanced cryptographic protocols that enable one party to prove the truth of a specific statement to another party without revealing any information beyond the validity of the statement itself (Tao et al., 2023). This technique ensures that the verifier learns nothing except that the statement is true, thereby protecting any underlying private data. This is very useful when privacy needs to be maintained during authentication or verification processes, as it eliminates the need to exchange or expose sensitive data directly.

By employing ZKPs in vehicle-to-anything (V2X) communications, a vehicle can authenticate or validate certain required conditions to other entities within the network without disclosing any additional, sensitive information. For instance, a vehicle could prove it has priority at an intersection (like an emergency vehicle might) without needing to reveal its exact location or any identifying details about its driver or passengers. ZKPs thus support various practical use cases. For example, consider a scenario where a vehicle needs to establish that it has the right to access a restricted traffic lane (such as a carpool lane) without revealing the number or identity of its occupants. Using ZKPs, the vehicle can simply provide proof that it meets the necessary criteria for the lane, such as having the minimum required number of passengers, without actually revealing who or how many passengers are inside. This method ensures compliance with traffic laws while upholding the privacy of the individuals within the vehicle.

Another, less-expected but impactful, use of ZKPs in the automotive sector involves financial transactions or subscription-based services tied to the vehicle. A vehicle could use ZKPs to prove that it has an active subscription for a particular service, e.g. to automated toll payments, without disclosing the details of the owner's account or personal information. This not only streamlines the transaction process but also enhances security by minimising the data exposed during transactions. As such, ZKPs essentially help mitigate privacy risks and allow to build trust in V2X communications.

3.3.6 Federated learning

This is a decentralised approach to machine learning that offers significant privacy advantages by enabling multiple participants (e.g., the connected vehicles themselves) to contribute to the development of a shared machine learning model without the need to share their individual data sets (Chellapandi et al., 2023). Instead, the learning algorithm is sent to each participant where it is trained locally on their data. After the training, only the updated model parameters or improvements, not the data itself, are sent back to a central server or aggregator. This means that the raw data generated by each participant stays on their device, reducing the risk of data breaches and exposure during transmission.

As each vehicle generates substantial amounts of data about its operations, environment, and driver behaviour, vehicle manufacturers and service providers can tap into this rich dataset by using federated learning to improve system-wide algorithms such as those used for automated driving, predictive maintenance, or traffic management, without compromising the privacy of the individual data sources.

The privacy benefits of federated learning are enhanced by the earlier mentioned techniques of differential privacy and encryption. Differential privacy can be applied during the training process on local devices to add noise to the model updates, thereby ensuring that these updates do not reveal specifics about the underlying data. Encryption ensures that any data transmitted, such as the model parameters, is secure against interception. These additional layers of privacy protection help mitigate any residual risks of information leakage during the federated learning process.

Moreover, federated learning not only improves privacy but also efficiency and scalability. It eliminates the need for a massive centralised data storage and processing infrastructure, reducing costs and potentially speeding up the learning process as data does not need to be transferred over the network. This scalability is particularly useful in the automotive industry where the number of connected devices (vehicles) is large and geographically dispersed.

In the context of connected vehicles, federated learning can be particularly transformative. For example, it could enable a fleet of vehicles to learn from collective experiences to improve safety features or optimise fuel consumption without ever sharing specific data about individual trips or driver behaviours. This collective learning capability could also extend to traffic pattern analyses where vehicles contribute to city-wide traffic management strategies without compromising the privacy of the drivers' location data.

3.3.7 Legal frameworks and regulations

In Europe, the privacy of data in connected vehicles is primarily governed under the GDPR. However, specific guidelines tailored to the nuances of connected vehicles have been developed by the European Data Protection Board (EDPB), as elaborated in (EDPB, 2020). They outline a set of comprehensive recommendations and regulatory frameworks for handling personal data within connected vehicle ecosystems (implementing data protection principles in the design phase of vehicle). Already adopted in 2021, they emphasise the importance of compliance with the GDPR and other relevant laws in the processing of such data. They detail the necessity of data minimisation, specifying that only essential data should be collected and processed. They outline several categories of data that can be processed, such as location data, biometric data, technical vehicle data, driving behaviour data, and personal data for infotainment systems. The guidelines advocate for transparent communication with data subjects about data handling practices and affirm the rights of individuals to access, rectify, and delete their data, thus ensuring users have significant control over their personal information. Additionally, the guidelines underscore the importance of robust security measures such as encryption and secure data transmission, and establish strict protocols for data sharing with third parties and international data transfers, ensuring compliance with legal standards. Special attention is given to various use cases, such as emergency data processing with systems like eCall, processing for mobility services, and scenarios involving employer-managed company cars, highlighting the legal bases for these activities.

The by now well-known ITS Directive 2010/40/EU, particularly in its revised form as of 2023, aims to ensure the coordinated deployment ITS across the European Union. This directive underscores the importance of privacy and data protection within the realm of connected vehicles. It extends to cover emerging services that involve significant data communication, like multimodal information systems, booking, ticketing, and automated mobility. Crucially, the directive focuses on creating a framework where data can be safely and effectively shared across different transport modes, ensuring interoperability and standardisation. It sets out specific requirements for digital communication between vehicles and infrastructure, emphasising the secure handling and privacy of the user data involved, which is integral to maintaining trust in the ITS ecosystem.

The draft C-ITS Delegated Act[7] focuses on enhancing the deployment and operational use of C-ITS across the European Union. Its primary goal is to foster a safe and efficient road transport network through the use of advanced communication technologies between vehicles and infrastructure. The act, similarly to the original ITS Directive, also emphasises the importance of interoperability and standardisation to ensure that vehicles can communicate seamlessly with each other and with traffic infrastructure. The act supports the use of both ITS-G5 and 5G technologies, considering them complementary rather than exclusive. This approach aims to leverage the immediate benefits of mature ITS-G5 technology for safety applications, while also remaining open to incorporating 5G as it becomes more viable for broader applications such as traffic management and more complex levels of automated driving. Overall, the C-ITS Delegated Act is more designed to prevent market fragmentation, thus ensuring a high level of network and information security across the C-ITS sector.

[7] See also https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=PI_COM%3AC%282019%291789 and https://clepa.eu/mediaroom/c-its-delegated-act-five-facts/.

In addition to this, the Security Policy for the Deployment and Operation of European Cooperative Intelligent Transport Systems focuses on establishing robust information security frameworks to protect the privacy and integrity of data in connected vehicles (EC, 2023). It mandates that C-ITS station operators implement certified information security management systems conforming to standards like ISO-27001[8]. These systems must cover all operational C-ITS stations and the data they process, ensuring confidentiality, integrity, and availability of the information. The policy emphasises risk management, requiring regular risk assessments to identify and mitigate potential threats to the system. C-ITS stations must classify information based on its impact on confidentiality, integrity, and availability, and use this classification to guide risk management processes. Specific controls are outlined for communication between C-ITS stations to safeguard data transfer, ensuring that any personal data transmitted maintains privacy in compliance with the GDPR.

Note though that, as remarked by (Berndt-Tolzmann et al., 2022), the integration of connected vehicles is a necessity for C-ITS. Road authorities and operators should develop strategic plans and allocate budgets tailored for C-ITS, enhance expertise in C-ITS technology, and engage in standardisation and testing to ensure systems compatibility and security. Specifically, adhering to IT security and privacy standards to support the safe deployment of connected vehicles remains on the forefront. By learning from impact evaluations of C-ITS pilots and sharing learned lessons, we can define robust governance and business models to facilitate the integration of C-ITS services.

The Safety-Related Traffic Information (SRTI) Directive, as part of the ITS Directive, does not explicitly focus on privacy for connected vehicles. However, it contributes indirectly by requiring the secure and compatible exchange of traffic and travel information across Europe. Unsurprisingly, by emphasising interoperability and standardization, the directive ensures that systems managing safety-related data do so in a manner that aligns with broader EU regulations on data protection and security. This supports the foundational privacy principles by mandating that any data exchange, including that which may involve personal information, is handled securely and in accordance with established European privacy standards.

The EU-Wide Real-Time Traffic Information (RTTI) Directive[9], formally known as the Commission Delegated Regulation (EU) 2022/670, outlines measures to ensure the privacy and data protection of users in the realm of connected vehicles and real-time traffic information. This regulation supplements Directive 2010/40/EU and specifies the types of data that can be collected, the purposes for which it can be used, and the obligations on data providers to protect this data. A significant aspect of this regulation is the emphasis on ensuring that data collection and processing is done in a way that respects the privacy of individuals.

[8] ISO-27001 outlines the requirements for establishing, implementing, maintaining, and continually improving an information security management system. It also includes requirements for the assessment and treatment of information security risks tailored to the needs of the organisation.

[9] https://transport.ec.europa.eu/transport-themes/intelligent-transport-systems/road/action-plan-and-directive/safety-related-traffic-information-srti-real-time-traffic-information-rtti_en

Additionally, the ISO/IEC 15408-2:2022 standard[10] (ISO/IEC. 2022) provides a detailed framework for evaluating IT security, with specific components applicable to connected vehicles. It outlines security functional components which can be crucial for ensuring privacy in connected vehicles, addressing the common security requirements and providing structured guidance for protecting user data and managing communication channels securely. This standard assists manufacturers and developers in embedding robust privacy controls into vehicle systems, helping safeguard against unauthorised data access and ensuring the integrity and confidentiality of the information exchanged between vehicles and networks. The standard identified four key attributes that relate to privacy, i.e. (i) anonymity, (ii) pseudonymity, (iii) unlinkability, and (iv) unobservability. Here, anonymity alone is insufficient for protection of an ITS user's privacy and unsuitable as a solution for ITS, as one of the main requirements of ITS is that the system should be observable in order to provide improved safety. Consequently, pseudonymity and unlinkability offer the appropriate protection of the privacy of a sender of basic ITS safety messages (cf. CAM and DENM). Pseudonymity ensures that a system may use a resource or service without disclosing its identity but can still be accountable for that use. Unlinkability ensures that a system may make multiple uses of resources or services without others being able to link them together.

As an example of a former Member State, the Privacy and Electronic Communications Regulations (PECR) in the UK sits alongside the GDPR, more specifically the Data Protection Act (DPA), and offers specific privacy rights concerning electronic communications. The regulations are particularly focused on the management of marketing calls, emails, cookies, and other forms of digital marketing based on electronic communications. While not specifically tailored for connected vehicles, PECR has implications for their operation, particularly in the context of user data transmitted via electronic communications. Under PECR, any use of cookies or similar technologies in connected vehicles that track location or user behaviour would require clear consent from users. This aspect is crucial when vehicles collect data to provide or enhance services, such as traffic information, route suggestions, or vehicle diagnostics that involve transmitting data over the internet. Additionally, if a connected vehicle's services involve any form of marketing communications via electronic channels, like promotional updates or service announcements, PECR mandates that these must be conducted with the user's prior consent, ensuring that privacy is maintained in accordance with user preferences and legal requirements. Thus, manufacturers and service providers must consider PECR when designing and implementing connected vehicle technologies that involve the processing of personal data through electronic communications.

[10] To be succeeded by the ISO/IEC WD 15408-2 standard which is currently in its preparatory phase.

Looking further outside Europe, we note that in the United States there is no specific federal regulation that comprehensively covers the privacy of connected vehicles. Instead, several state laws may apply, and federal guidelines have been proposed. The National Highway Traffic Safety Administration (NHTSA) has issued guidelines to protect consumer privacy in connected vehicle technologies. These emphasise transparency, choice, respect for context, data minimisation and de-identification, data security, integrity and access, and accountability. Continuing, China has been advancing regulations that focus on data security and privacy for connected vehicles under its broader push for strengthened data protection laws. The Cybersecurity Law, implemented in 2017, and the Data Security Law, effective from 2021, both provide a regulatory framework that impacts connected vehicles, especially concerning the collection and handling of personal data. Finally, Japan has also taken steps towards regulating connected car data under its Personal Information Protection Commission. This provides guidelines that direct how personal information should be handled, ensuring user data collected by vehicles is protected under Japan's privacy laws.

From the perspective of the vehicle manufacturers (i.e. the OEMs, Original Equipment Manufacturers), the growing regulatory focus on data privacy in connected vehicles presents both challenges and opportunities. Many OEMs believe that stringent GDPR and EDPB guidelines ensure consumer trust, which is essential for the wide-scale adoption of connected technologies. However, these regulations also require significant adjustments to vehicle design, manufacturing processes, and data management systems. Manufacturers are concerned about the costs of implementing advanced privacy-by-design frameworks, including pseudonymisation, encryption, and secure communication protocols (Raes et al., 2020).

Furthermore, OEMs emphasise the need for clarity and standardisation in regulations, as inconsistent guidelines across jurisdictions can lead to fragmentation in the market. They argue that interoperability and a clear regulatory framework are vital to avoid hindering innovation in V2X technologies. OEMs also support a balanced approach, where regulations allow for sufficient data collection to enhance vehicle safety and performance while maintaining user privacy. They recognise that compliance with data minimisation principles can sometimes limit the scope of innovation, particularly in areas like automated driving, which relies heavily on large datasets.

3.4 State of practice of mitigation measures

In the following sections we first give some background against which the current mitigation measures are considered, followed by several examples of currently used and usable techniques in different domains and scenarios that are applicable for connected vehicles.

3.4.1 Context

A decade ago, (de Montjoye et al., 2013) raised privacy concerns stemming primarily from the extensive data connected vehicles can collect, ranging from location histories and driving patterns to even conversations inside the vehicle. This data collection would raise critical issues regarding driver consent, as drivers must be fully aware and agree to what data is collected. Moreover, the potential for data breaches poses significant risks, as connected vehicles effectively function as mobile data centres. The usage and sharing of vehicle data also present major privacy challenges. For instance, the authors indicated that data could be used by insurance companies to alter premiums or by advertisers to target ads based on detailed behavioural profiles[11]. Additionally, the capability of connected vehicles to facilitate surveillance and tracking by both government and private entities was and is a concerning prospect, requiring strict regulations to prevent abuse. Concluding, they underscore the need for robust legal frameworks and advanced technological measures, like encryption and anonymisation techniques, to safeguard individual privacy rights effectively without stifling innovation.

As the European Commission further looked at the ethics of connected and automated vehicles, they emphasised the importance of privacy in the context of the extensive data collection involved (EC, 2020b). They stress adherence to the GDPR, advocating for data minimisation and user consent for non-essential data uses. Key recommendations include safeguarding informational privacy by requiring explicit user consent for data usage beyond vehicle operation, providing mechanisms for users to control their data (such as rights to access, rectify, and erase data), and enhancing transparency to empower users fully. Reiterating this point in their data strategy (EC, 2020a), the European Commission makes a case for making non-personal and industrial data available and usable, with the need for proper data governance, and the establishment of a single European data space. In this context, while recognising the generation of vast amounts of data by connected vehicles useful for various innovative mobility-related services, they nevertheless emphasise the secure management of sharing and access to in-vehicle data in compliance with data protection rules and competition laws, in such a way that competition is maintained and that multiple players can innovate and provide services.

[11] The irony here is that, in the meantime, this is actually happening but in a subtly different context: drivers who drive more cautiously (and are also monitored) pay lower insurance fees, leading to less accidents on the road, implying a win-win for both the drivers and the insurance company.

Aside from the technical considerations, it is also necessary to take privacy perceptions and decisions of users into account, as done, e.g., by (Cai and Xiong, 2023). They conducted an extensive study involving just shy of 600 participants to assess how privacy concerns vary across different V2X scenarios such as cooperative autonomous driving, road safety, traffic management, and infotainment applications. They found that participants generally perceive greater benefits and fewer privacy risks in scenarios where data sharing is essential for operational safety and efficiency. Furthermore, the study revealed that priming users with privacy risk information influenced their willingness to share data, although this effect varied depending on the user's prior experience with vehicle connectivity and assistance systems. Additionally, the research delved into the technical and behavioural aspects influencing user decisions around data privacy. It highlighted a privacy-safety trade-off, where users might overlook privacy concerns for perceived safety benefits. Factors like misconceptions about data collection and use, as well as the novelty of CAV technologies, significantly shaped users' decisions. The findings suggested a complex interplay between perceived benefits, privacy risks, and user experience in shaping attitudes towards data sharing.

In a nutshell, the previous sections have shown that there are several pressing and recurring matters to consider when discussing mitigation measures. In summary, (Rebiger et al., 2019) provide a concise overview of the most relevant aspects in this respect:

- **Lack of transparency**: the complexity of data processing in connected cars makes it difficult to inform users clearly about what data is collected, by whom, and for what purpose. This complexity challenges the enforcement of privacy policies and user consent protocols.
- **Excessive data collection**: there is a risk that the vast amount of sensors and data collection points in connected cars could lead to unnecessary collection of personal data, not strictly required for the provided services.
- **Data retention**: proper data retention policies are crucial as there is a risk that data could be stored longer than necessary, increasing the risk of misuse or unauthorised access.
- **Control over personal data**: users often lack sufficient controls to manage their personal data effectively within connected car systems, which complicates the ability to maintain privacy.
- **Purpose limitation**: data collected for specific purposes, like vehicle maintenance, could be repurposed for other uses such as insurance adjustments or law enforcement surveillance without clear user consent.
- **Security risks**: as part of IoT, connected cars are susceptible to various security risks including cyberattacks, which could compromise both personal data and vehicle operation.

Thus, the pervasive data collection in connected cars raises significant privacy issues, especially since they can generate up to many gigabytes of data per hour, much of which is personal data. This data can include biometric, health, location, and communication details, raising concerns about how it is used and protected. The need to adhere to data protection regulations such as GDPR, implementing robust security measures, and developing clear guidelines for data access and control in connected cars is quite apparent. These measures are vital to mitigate privacy risks and enhance user trust.

This latter points is further made for systems that exchange real-time data between vehicles and infrastructure. They process personal data by broadcasting continuous messages containing vehicle specifics and kinematic data, thus raising significant privacy risks. These data exchanges qualify as personal data due to identifiable details linked to each vehicle. Therefore, stringent privacy safeguards are needed, embedded within a robust legal framework under GDPR, taking measures such as data minimisation techniques and secure message broadcasting practices to protect individuals' privacy (EC, 2017). As such, there is a need for EU-wide legislative action to ensure that the deployment and operation of connected vehicles within C-ITS adhere to data protection laws, ensuring the processing is lawful, necessary, and proportionate while maintaining the public trust and safety objectives of the technology.

3.4.2 Dealing with GPS traces

(Kamola, 2015) explored a new methodology for anonymising GPS data in a way that supported traffic analysis while enhancing privacy. It shifted from traditional individual trace anonymisation to anonymising the whole road graph, where GPS locations are projected as distances from road intersections, abstracting physical node locations. This technique not only preserves the utility of the data for monitoring and analysing traffic flows, behaviours at intersections, and other driver behaviours essential for managing connected vehicle ecosystems, but also enhances privacy by detaching the data from specific geographic details. Their work demonstrated the approach using real traffic data, thereby showcasing its potential to maintain robust traffic analysis and behavioural studies capabilities without compromising individual privacy.

An example of this technique would be to apply it to the GPS traces (from the on-board unit) supplied by ViaPass in Belgium. To transform these GPS data points into relative distances from the nearest road intersections or junctions, we need to first map these locations onto a pre-anonymised road graph that represents the road network as a series of nodes (i.e. intersections or junctions) and edges (i.e. the road segments between nodes). Each GPS point is then projected onto the closest edge in this graph. By calculating the distance from the nearest node along this edge and expressing each GPS point as a one-dimensional position relative to this node, this projection effectively converts two-dimensional GPS coordinates into a linear measurement that retains essential data for route analysis (and possibly even toll calculation), without revealing the exact geographic locations. The resulting dataset comprises distances from nodes rather than specific coordinates, significantly enhancing privacy while maintaining the data's utility for traffic and logistic analyses.

(Maouche, 2019) focused on addressing the vulnerabilities in location privacy protection mechanisms against re-identification attacks, which aim to match anonymised location data with identifiable users, thus compromising their privacy. The central theme revolved around the concept that the conventional techniques (like simply removing user IDs or obscuring certain data points) are insufficient as the mobility data itself can serve as a quasi-identifier due to its unique patterns. In his research, he introduced novel methodologies for both simulating attacks to evaluate the robustness of these techniques as well as developed more effective countermeasures. This is an evolving field, as with the rise of more powerful machine learning techniques, new re-identification attacks exploit weaknesses in existing privacy protections by examining the spatial and temporal patterns in mobility data to associate anonymous traces back to known users. To counteract these vulnerabilities, a new privacy mechanism was proposed that was designed to enhance user anonymity by altering mobility traces in a controlled manner. By modifying a user’s mobility trace to resemble that of another user, the unique spatial-temporal signature that could otherwise be exploited by attackers was disrupted. This method not only complicated the task of re-identification but also considered the utility of the data, attempting to maintain its value for legitimate applications while safeguarding user privacy.

3.4.3 Dealing with intersection control

(Tan and Yang, 2024) presented a method for privacy-preserving adaptive traffic signal control in environments with connected vehicles. Their research addressed the privacy risks associated with using detailed data stemming from these vehicles, such as real-time trajectories and personal preferences. While they are valuable for optimising urban traffic systems, they can expose sensitive information about individuals. By integrating secure multi-party computation (see Section 3.3.4) and differential privacy (see Section 3.3.2), the authors proposed a system that aggregates the data without requiring vehicles to reveal their private information directly, thus maintaining privacy while still allowing effective traffic management.

Their developed system utilises a linear optimisation model for traffic signal control based on securely aggregated traffic data. This model was designed to minimise traffic delays and queue lengths at intersections, adapting in real time to changing traffic conditions while ensuring that individual vehicle data remains confidential. A key feature was the use of stochastic programming to accommodate the uncertainty and noise introduced by the privacy-preserving mechanisms, notably differential privacy, which added random noise to the data to prevent individual vehicles' data from being identified.

Empirical results demonstrated that the proposed methods effectively balanced privacy and utility, meaning they maintained high operational performance in traffic signal control while protecting individual privacy. The system showed potential for implementation in real-world traffic management systems where adoption rates of connected vehicles are increasing, highlighting the feasibility of privacy-preserving approaches in critical infrastructure.

3.4.4 Application of European regulations

3.4.4.1 Privacy-centric data handling

The PoliVisu (European) project provided interesting applications for handling and anonymisation of privacy-sensitive data to support policy-making without infringing on individual privacy (Raes et al., 2020). It covers various techniques for data protection, focusing on European privacy regulations, particularly the GDPR, and applies these to different types of transport-related data used in smart city contexts. It looked at several data types, including automated number-plate recognition data, crowd-sourced traffic counts, floating car data, cellular and Wi-Fi sniffing data, and road accident data. Each data type was explored in detail, discussing how data was collected, processed, anonymised, and the specific challenges encountered in ensuring privacy compliance. Real-world applications were provided by performing the different data handling techniques in a set of pilot cases conducted in various European cities. The project stressed the importance of adopting a precautionary principle to protect the privacy and anonymity of citizens[12], while acknowledging the potential of using personal data for public benefits in policy-making. A central role in all this was the concept of privacy by design, whereby privacy measures are not merely add-ons but integral components of the development process, ensuring that all data collected from the vehicles is handled transparently and with respect to user consent, thus maintaining trust and safeguarding against potential breaches. Their mitigation measures relied heavily on data anonymisation, data pseudonymisation, minimal data collection and data aggregation (see Section 3.3.1), as well as limiting data retention to ensure that data is only stored for as long as necessary for the intended purpose and securely deleting it afterward to reduce the risk of misuse. In addition they employed secure hashing algorithms, such as PBKDF2[13], which is resistant to brute-force attacks and provides effective encryption of identifiers (like number plates), and introducing randomness in data collection and reporting intervals (see Section 3.3.2) to prevent forming recognisable patterns.

3.4.4.2 Relevant European regulations

In this following, we will explore several key regulatory frameworks and initiatives shaping the European digital landscape. Our focus will be on the European Data Governance Act, which aims to facilitate data sharing across the EU, the European Data Act, designed to regulate data access and usage, the European e-Privacy Directive, which governs electronic communications and privacy, and the European AI Act. Additionally, we will delve into relevant recommendation from the C-Roads initiative, which is a cooperative ITS deployment project. Together, these elements provide a cornerstone of the EU's approach to data management, privacy, and intelligent transport systems.

[12] In enforcement cases, data used for issuing fines (e.g., ANPR data) is typically pseudonymised rather than fully anonymised. This ensures that personal information is protected during data processing, but the data can be re-identified when necessary, such as for legal enforcement actions. This approach balances compliance with privacy regulations like the GDPR while still allowing authorities to access identifying details when issuing fines.

[13] PBKDF2 (Password-Based Key Derivation Function 2) is a cryptographic algorithm used to derive a secure encryption key from a password. It enhances security by applying a hashing algorithm multiple times to a password along with a salt, significantly increasing the difficulty for attackers to perform brute force attacks.

3.4.4.2.1 Data Governance Act

The EC's Data Governance Act[14] (DGA) is a legislative framework aimed at fostering the availability of data for use in the European Union, enhancing trust in data sharing, and establishing mechanisms for data governance. Adopted in 2021, the DGA addresses the challenges posed by the exponential growth of data and the need for clear rules and standards to facilitate data access and reuse, especially for public sector data, personal data, and data held by private entities. The DGA is part of the broader European Data Strategy, which aims to create a single market for data and ensure Europe's global competitiveness.

One of the core elements of the DGA is the establishment of common European data spaces in strategic sectors such as health, environment, energy, agriculture, and finance. The Act sets out clear rules for the reuse of certain categories of protected data held by public sector bodies, aiming to balance data accessibility with the protection of sensitive information.

The DGA introduces the concept of data intermediation services, which act as neutral data brokers between data holders and data users (e.g., C-ITS services that rely heavily on data sharing between vehicles, infrastructure, and service providers). These intermediaries are designed to facilitate voluntary data sharing while ensuring compliance with data protection regulations and fostering trust among participants. The Act requires these service providers to be registered and adhere to a set of requirements to ensure transparency, impartiality, and security in their operations. This helps to mitigate risks associated with vehicle re-identification by ensuring that only authorised entities have access to sensitive data and that such data is anonymised or pseudonymised appropriately to protect individual privacy. These services are required to adhere to stringent regulations to maintain neutrality, transparency, and security. By ensuring that data intermediaries operate under clear and enforceable rules, the DGA reduces the risk of unauthorised data access and misuse, thus protecting the privacy of individuals using C-ITS services and preventing vehicle re-identification. The DGA's provisions on the reuse of public sector data include mechanisms to ensure that such data is shared in a way that respects privacy and confidentiality. For C-ITS, this means that data collected by public authorities, such as traffic data or road usage statistics, can be made available for innovation and service improvement while ensuring that individual vehicles cannot be re-identified.

Additionally, the DGA establishes a framework for data altruism, encouraging individuals and organisations to voluntarily make their data available for the common good, such as for scientific research or public interest projects. The Act provides for the creation of national registers for data altruism organisations and sets out guidelines for ensuring that data is shared ethically and with the necessary consent from data subjects. In the context of C-ITS, this can support research and development of new technologies and services aimed at improving transportation safety and efficiency. However, it also places a strong emphasis on obtaining explicit consent from data subjects and ensuring that data shared altruistically is used ethically and with appropriate safeguards against re-identification.

[14] Formally known as Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act).

As the DGA includes provisions for enhancing data portability and interoperability, emphasising the importance of developing technical standards and protocols to ensure seamless data exchange and integration across the EU.

3.4.4.2.2 Data Act

The EC's Data Act[15] is another significant legislative measure aimed at fostering a robust data economy within the EU. This act complements the DGA by focusing more on the rights and obligations surrounding data access and use, particularly concerning data generated by devices and related services, thereby trying to balance enabling data-driven innovation while protecting individual rights.

The Data Act establishes rules for fair access to and use of data generated by Internet of Things (IoT) devices, which are integral to C-ITS services. The aim is that data generated by vehicles and infrastructure can be shared among different stakeholders under clear and fair conditions. The act mandates that users have control over who can access their data, thus enhancing privacy protections. To this end, it includes provisions for technical measures such as anonymisation and data minimisation. And similar to the DGA, the Data Act also emphasises the importance of interoperability and standardisation in data sharing.

Users, such as vehicle owners, have the right to access data generated by their devices (e.g., sensors for telematics data), transfer it to other service providers (e.g., fleet management systems), and share it with authorised third parties (e.g., for services like navigation or maintenance) under fair and non-discriminatory conditions. On the obligations side, manufacturers and service providers must provide data access to users and their authorised third parties, ensuring the terms are fair and reasonable. Here, the requirements for anonymisation and pseudonymisation play a crucial role.

3.4.4.2.3 ePrivacy Directive

The ePrivacy Directive[16] provides legislation in the European Union dealing with privacy and electronic communications. It complements the GDPR by specifically addressing the confidentiality of communications and the processing of personal data in the electronic communications sector.

It does this by prohibiting the interception and surveillance of communications without the mandatory consent of the user. In the context of C-ITS, this means that any data (particularly tracking and location data) transmitted between vehicles, infrastructure, and service providers must be kept confidential and protected against unauthorised access.

[15] Formally known as Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act).

[16] Formally known as Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications).

Furthermore, the directive promotes data minimisation and anonymisation, including the implementation of appropriate security measures by service providers in order to protect the personal data they process. Practically this boils down to securing the data transmission channels, storage systems, and any other points where data might be vulnerable to unauthorised access or breaches. Seemingly a bit less related, the directive also regulates the use of cookies and other tracking technologies, which are often used to collect data about users' online behaviour. However, even for C-ITS services similar technologies might be used to track vehicle movements and behaviour.

3.4.4.2.4 AI Act

The European AI Act[17] (EAA) has relevance to privacy concerns for C-ITS services due to its risk-based classification of AI systems. The act distinguishes between minimal risks, limited risks (of AI systems with specific transparency obligations), high risks, and unacceptable risks. AI systems in these contexts could be deemed high-risk, given their significant impact on safety and personal data. As a result, they are subject to stringent requirements, including robust data governance and transparency measures. These requirements directly address privacy concerns by ensuring that personal data is managed securely and ethically, focusing on data accuracy, minimisation, and purpose limitation.

Transparency and accountability are central to the EAA, compelling manufacturers and operators of AI systems in connected vehicles to provide clear information about personal data collection, processing, and usage. Additionally, the EAA emphasises human oversight and the robustness of AI system design to prevent unintended consequences.

The EAA also mandates conformity assessments to ensure compliance, requiring manufacturers and service providers to demonstrate that privacy-by-design principles are integrated into AI systems' development and deployment. This ongoing compliance ensures that privacy regulations are met, providing continuous protection for personal data in connected vehicles. Furthermore, the EAA's requirement for post-market monitoring (i.e. the ongoing surveillance and assessment of AI systems after they have been released into the market) and incident reporting ensures that any privacy issues are promptly identified and addressed.

Finally, the EAA complements existing privacy frameworks, such as the GDPR, by providing additional AI-specific safeguards. This alignment enhances protection measures for personal data, ensuring that AI systems in C-ITS services and connected vehicles comply with both GDPR and EAA requirements.

[17] Formally known as the European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)).

3.4.4.2.5 C-Roads recommendations

The C-Roads Platform[18] is a joint initiative of European Member States and road operators to implement and harmonise C-ITS across Europe. The goal is to ensure the interoperability of C-ITS services while supporting the deployment of cross-border C-ITS infrastructure. To this end, the C-Roads Platform has published a roadmap (C-Roads Platform, 2024): this is not a typical roadmap that projects the readiness of future specifications. All C-ROADS use cases in the published profiles have been tested and validated in pilot deployments in the real-world before they were published. Having such an implementation and having the opportunity of feedback from cross-border tests ensures that the implementation is feasible, working and thus ready to market.

The primary goal of the C-Roads Platform is to ensure that C-ITS services are interoperable across different countries and regions in Europe. This involves defining common standards and specifications for C-ITS deployment. By promoting interoperability, the C-Roads Platform ensures that privacy-preserving techniques and data protection measures are consistently applied across borders, reducing the risk of vehicle re-identification when data is shared between different systems and countries.

The platform establishes common security policies to protect data integrity and privacy in C-ITS communications, including encryption and authentication methods to secure data exchanges between vehicles and infrastructure. These common security policies ensure that all participating entities adhere to high standards of data protection. To this end, the Platform also provides guidelines and best practices for data protection in C-ITS services, aligning with EU data protection laws such as GDPR. These guidelines emphasise the need for anonymisation and pseudonymisation.

Additionally, the platform supports pilot projects and field tests to validate C-ITS services in real-world conditions, providing valuable insights into practical privacy challenges and allowing the development of more robust privacy-preserving techniques.

Furthermore, the C-ROADS Steering Committee's Working Group 1 (C-Roads Platform, 2021) focused on the data protection issues associated with delivering interoperable C-ITS messages across Europe. The primary concern was the protection of driver identity within vehicles, managed through pseudonym certificates in the European PKI system. However, data protection authorities have expressed concerns regarding the traceability of vehicles even with this technology. The group's efforts enhanced awareness and common understanding of the GDPR related to C-ITS services among road operators and other stakeholders. The group gathered and disseminated information from various C-Roads projects to address GDPR compliance in C-ITS. This revealed significant disparities in knowledge about data protection, after which they identified key GDPR issues such as the legal status of authentication certificates and the unencrypted broadcast of short-range communications from vehicles. They suggested measures such as default 'receive only' settings, data protection by design, and the avoidance of centralised databases of exchanged messages to mitigate these concerns. Nevertheless, they also emphasised the importance of further assessment and potential new legislation to support the full potential of C-ITS, considering the balance between road safety benefits and data privacy requirements.

[18] https://www.c-roads.eu/

3.4.4.2.6 Other relevant European regulations

In addition to the previously mentioned regulations, and aside from the GDPR, ITS Directive, and the latter's supplemental, there are several other directives and acts important and relevant, forming a comprehensive legal framework that supports the secure, efficient, and privacy-respecting implementation of C-ITS services across the European Union. We briefly highlight them and their relevance for C-ITS in the following paragraphs.

- **Directive on Security of Network and Information Systems (NIS Directive)**
  The NIS Directive[19] is one of the first pieces of EU-wide legislation on cybersecurity. Its goal is to achieve a high common level of security of network and information systems across the EU. By mandating robust cybersecurity measures for network and information systems, such as multi-factor authentication (MFA) and end-to-end encryption (E2EE), the NIS Directive helps prevent unauthorised access to C-ITS data, thereby protecting vehicle identities and preserving user privacy.

- **Free Flow of Non-Personal Data Regulation**
  This regulation[20] wants to remove obstacles to the free movement of non-personal data within the EU, thereby creating a more competitive and integrated data market. It complements the GDPR by ensuring that data, which does not contain personal information (e.g., by anonymising and aggregating traffic flow data), can move freely across borders, and for C-ITS data to be used and shared for traffic management and analytics.

- **eIDAS Regulation**
  The eIDAS Regulation[21] establishes a framework for electronic identification and trust services for electronic transactions in the internal market. It enhances trust in electronic transactions and is crucial for the secure and efficient functioning of C-ITS. The regulation ensures that only authorised entities can access C-ITS data, e.g., by using digital certificates (like PKI) to authenticate vehicles and infrastructure elements.

- **Public Sector Information (PSI) Directive**
  This directive[22] encourages the re-use of public sector information for private or commercial purposes, promoting transparency and the creation of value-added services. This is relevant for C-ITS, which can leverage publicly available anonymised data for better traffic management and other services without exposing personal vehicle data, e.g., using anonymised traffic sensor data to optimise traffic light timings to enhance traffic flows.

[19] Formally known as Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union.
[20] Formally known as Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union.
[21] Formally known as Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC.
[22] Formally known as Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information.

- **European Electronic Communications Code (EECC)**
  The EECC[23] consolidates and updates the EU's regulatory framework for electronic communications. It aims to ensure a competitive market, improve digital infrastructure, and enhance consumer protection, which is relevant for the communication aspects of C-ITS. The EECC protects the integrity of data transmitted in C-ITS, preventing interception and unauthorised access, e.g., by making use of secure E2EE cryptographic communication protocols like Transport Layer Security (TLS) for all C-ITS data exchanges ensures data is encrypted during transmission, safeguarding against eavesdropping and unauthorised access.

- **The Knowledge Base on Connected and Automated Driving[24]** is not a regulation but branched out of the Horizon 2020 ARCADE Support Action and is currently maintained by the Horizon Europe FAME project in line with the European Partnership on CCAM. Regarding data-protection recommendations they emphasise the importance of creating trust between data providers, data owners, and data consumers. The data provider is responsible for ensuring that data are handled according to agreements, contracts, and the relevant legal context. This trust is built by ensuring that the data consumer has robust data protection procedures, which are documented and proven effective. The guidelines apply to various scenarios where data is shared between organisations, necessitating discussions on data categories, risks, access methods, purposes of data exchange, security requirements, and legal compliance. They knowledge base also outlines specific measures for data consumers, who must document their data-protection implementations before accessing data. This documentation should include a comprehensive overview of data usage plans, legal analyses for compliance with GDPR and national laws, and details on data protection infrastructure. Additionally, it should cover incident response plans, internal routines, personnel training, and mechanisms to prevent unauthorised access. In addition, the initiative also addresses the roles of different stakeholders in the data-sharing process, thereby highlighting the need for clear agreements and robust data governance procedures. Relevant recommendations include establishing physical and logical security requirements, defining data retention and erasure policies, and ensuring proper documentation.

[23] Formally known as Directive (EU) 2018/1972 of the European Parliament and of the Council of 11 December 2018 establishing the European Electronic Communications Code.
[24] https://www.connectedautomateddriving.eu/

3.4.4.3 Summary

While the previously listed regulations typically do not prescribe specific technical implementations, or concrete recommendations or solutions, they do create a (legal) framework for achieving a high common level of network and information system security across the EU. As they focus more on outcomes and principles, they consequently allow Member States and relevant entities themselves to determine the best technical measures to meet these requirements.

Diving deeper into the specific implications of these regulations for C-ITS, they typically refer back to techniques include anonymisation, pseudonymisation, data minimisation, differential privacy, encryption, secure multi-party computation, etc., as already elaborated upon in Section 3.3. It is interesting to notice that in other fields, such as healthcare, the same principles play are relevant for C-ITS communications to ensure the secure transmission of V2X data, protecting against unauthorised access and data breaches, and maintaining the privacy of individual drivers (Nelson, 2015).

3.5 Summary

This chapter provided a comprehensive overview of the current state of research and methodologies related to the re-identification of anonymised data in the context of connected vehicles. We began with a background on the significance of privacy in connected vehicle systems, emphasising how the data transmitted between vehicles and infrastructure, considered personal due to the inclusion of authorisation certificates and detailed location data, can be susceptible to deanonymisation. We highlighted the inherent privacy risks associated with C-ITS messages, which can reveal sensitive information about vehicle locations and user behaviours even when rudimentary anonymisation techniques are applied.

We then looked into various re-identification methods documented in literature, illustrating the vulnerabilities of anonymised location data. We discussed how minimal location data points can uniquely identify individuals within large datasets, a concept explored through case studies like the analysis of mobility traces from a European mobile operator. The methods to counter these threats were also reviewed, including suppression and generalisation strategies that enhance privacy protections. We further explored trends in vehicle re-identification technologies, such as the integration of multiple data modalities and the use of deep learning techniques, which improve identification accuracy but also pose new challenges for privacy.

Finally, we considered the state of the art of mitigation measures against deanonymisation. To this end, we covered a range of technical measures like differential privacy, synthetic data generation, and encryption techniques, alongside legal frameworks aimed at strengthening data protection. Additionally, practical implementations such as vehicle appearance modification and controlled disclosure of location data were discussed. In conclusion, we underscored the importance of a balanced approach that combines technological innovation with robust privacy safeguards, advocating for collaboration among governments, industry stakeholders, and privacy advocates to ensure ethical and responsible use of re-identification technologies in connected vehicle systems.

4 Impact study

4.1 Background

Following the insights obtained in the previous section, we now focus our attention on the next two points:

(1) What (types of) information about road users could be leaked from C-ITS data?

(2) What is consequently the potential impact on the data subject itself?

For (1) we will start in Section 4.2 by first looking at the specific content that is transmitted in the C-ITS data (i.e. what are the headers that are required per message type? which ones are optional? what message types are sent? what is the frequency with which the information is sent? etc.).

Based on the information provided and their key features, we can check for possible links with personal data. We should not dismiss any attributes or characteristics, even if they seem unlikely. This is because what seems hard or unlikely now might not be in the near future. As local processing power increases, re-identification could become an issue, even if it doesn't seem urgent now.

Closely related to the previously described work, we will also investigate in Section 4.4 what and how great the impact of such data breaches are on a re-identified individual. As it may be possible to detect behavioural patterns, implicit sensitive information, or even other properties of such individuals, this research may also provide us with an idea of which part of the road users that can be re-identified on the basis of leaked location data, and what the required efforts are to accomplish this.

4.2 Information that can be leaked from C-ITS data

In order to determine the breadth of information that can be leaked from C-ITS data, we will first consider the specific content that is transmitted by means of the types op messages that are sent, their frequency, and what headers and other relevant information is contained in them.

4.2.1 Types of V2X messages

V2X (C-ITS) message types (sets) and their subtypes are closely related to the different types of services that are foreseen. In Table 2 we present some of the relevant message sets for V2X communications, currently already standardised or in the process of being standardised (C2C-CC, 2018, Rondinone and Correa, 2018, and C-Roads Platform, 2023).

Table 2: Overview of some of the C-ITS message sets relevant for V2X communications.

Message	Description	Frequency	Main contents
BSM	Basic Safety Message	10 Hz (typically fixed)	Provides vehicle location, speed, heading, and other critical information for collision avoidance and traffic management
CAM	Cooperative Awareness Message	1-10 Hz	Transmits status information about a vehicle or road user to nearby vehicles and infrastructure
CPM	Collective Perception Message	1-10 Hz	Shares detected object information from a vehicle or infrastructure sensor systems to other vehicles and infrastructure
DENM	Decentralised Environmental Notification Message	Event-driven	Alerts nearby vehicles and infrastructure to hazardous events or conditions, e.g., weather, traffic jams, road works, etc.
IVIM	In-Vehicle Information Message	Event-driven or periodic	Provides in-vehicle signage and information, e.g., speed limits and warnings
MAP	Map Message	1 Hz or lower	Detailed road and intersection layout (cf. road and lane topology service and traffic light manoeuvre service)
MAPEM	Map Data Extended Message		Extends the MAP messages with additional information relevant to specific use cases
MCDM	Multimedia Content Dissemination Message	Event-driven or periodic	Distributes multimedia content to nearby vehicles or infrastructure
MCM	Manoeuvre Coordination Message	10 Hz	Coordinates manoeuvres between vehicles, such as lane changes or merging
PSM	Personal Safety Message	1-10 Hz	Broadcasts information about vulnerable road users (e.g., pedestrians, cyclists) to nearby vehicles and infrastructure
SPAT	Signal Phase and Timing Message	1-10 Hz	Provides information about the current and future status of traffic signals
SPATEM	Signal Phase and Timing Extended Message		Extends the SPAT messages with additional information relevant to specific use cases

SREM	Signal Request Extended Message	Event-driven	Allows vehicles to request signal priority or pre-emption (cf. traffic light control service)
SSEM	Signal request Status Extended Message	Event-driven	Provides status updates on signal requests, such as whether the request was granted

Note that in addition to these, there are also a range of non-standardised messages which we will not treat here. Examples of these can be found in (Rondinone and Correa, 2018), e.g., Lane Advice Message (LAM), i-GAME Cooperative Lane Change Message (iLAM), Cooperative Lane Change Message (CLCM), Convoy Management Message (CMM), Cooperative Sensing Message (CSM), Cooperative Speed Advising Message (CSAM), etc. They are typically defined in the context of (ongoing) research projects.

4.2.2 Structure of V2X messages

V2X messages typically contain a high-level structure that follows standardised formats to ensure interoperability. Any such logical grouping of related data elements within a message is referred to as a container. They help to organise the data in a structured and modular way, making it easier to interpret and process the information. Examples of these are:

- (i) a **header**: message identification, protocol version, message length, etc.
- (ii) a **payload**: the actual message content such as vehicle state data, etc.
- (iii) **metadata**: e.g., timestamps, sender identification, geographical information, etc.
- (iv) **security and integrity data**: digital signatures to ensure the authenticity and integrity of the message, along with extra encryption information
- (v) **optional extensions**: e.g., custom data fields for application-specific data, error correction information, etc.

In Table 3 we present the most relevant headers for each of the listed message types.

Table 3: Overview of some of the most relevant headers for the identified C-ITS message sets.

Message	Most relevant headers
BSM	Message ID, Message Count, Temporary ID, DSecond (time of the message), Latitude, Longitude, Elevation, Positional Accuracy, Transmission State, Speed Heading, Steering Angle, Acceleration Set, Brake System Status, Vehicle Size
CAM	Message ID, Station ID, Generation Time, Station Type, Latitude, Longitude, Altitude, Heading, Speed, Drive Direction, Vehicle Length, Vehicle Width, Longitudinal Acceleration, Yaw Rate, Acceleration Control
CPM	Message ID, Generation Time, Station ID, Station Type, Latitude, Longitude, Altitude, Speed, Heading, Perceived Object List (with details like object ID, type, position, speed, etc.)

DENM	Message ID, Station ID, Detection Time, Reference Time, Latitude, Longitude, Altitude, Validity Duration, Station Type, (Sub) Cause Code, Event History
IVIM	Message ID, Information Type, Information Content, Start Time, Duration, Latitude, Longitude, Station ID
MAP	Message ID, Intersection ID, Intersection Name, Latitude, Longitude, Elevation, Lane List (with details like lane ID, type, ingress/egress, etc.)
MAPEM	Message ID, Intersection ID, Intersection Name, Latitude, Longitude, Elevation, Detailed Lane List (with additional attributes compared to standard MAP)
MCDM	Message ID, Content Type, Content Length, Content Data, Source ID, Destination ID
MCM	Message ID, Generation Time, Station ID, Manoeuvre Type, Start Time, Duration, Intended Path (waypoints, speed, heading, etc.)
PSM	Message ID, Temporary ID, Generation Time, Latitude, Longitude, Speed, Heading, Acceleration, Type (e.g., pedestrian, cyclist, etc.)
SPAT	Message ID, Intersection ID, Intersection Name, Current Time, Signal Phase State, Time To Change, Intersection Status, Latitude, Longitude, Elevation
SPATEM	Message ID, Intersection ID, Intersection Name, Detailed Phase Timing Information (e.g., min/max time to change, pedestrian phases, etc.)
SREM	Message ID, Request ID, Station ID, Request Time, Requested Phase, Vehicle Information (type, speed, location)
SSEM	Message ID, Response ID, Request ID, Status, Intersection ID, Estimated Time of Service

From the table we can see that there exist common headers across all the listed communication messages, i.e.:

- **Message ID**: Every message type includes a unique identifier to distinguish it from other messages.
- **Station ID**: Most messages[25] include a unique identifier for the sending or receiving station (vehicle, roadside unit, etc.).

Additionally, the following header, while not universally present in every single message type, is quite common in many of them:

- **Generation Time**: To indicate the time when the message was generated, thereby providing a timestamp for the message.

[25] Note that for Personal Safety Messages and some other specific messages, this might be referred to as a Temporary ID or another specific ID type.

These common headers ensure that each message can be uniquely identified, traced back to its source, and properly time-stamped for effective communication and processing.

Regarding the Station IDs, we note that items such as Vehicle Identification Numbers (VIN) are generally not transmitted due to privacy and security concerns (e.g., a VIN can uniquely identify a vehicle and its owner). Instead, C-ITS messages use temporary and pseudonymous identifiers to protect the privacy of the vehicle and its occupants. These identifiers are designed to provide a level of anonymity and are frequently changed to prevent tracking of individual vehicles. Note that in certain specialised or controlled environments, such as fleet management or within secured systems (e.g., within a specific organisation or for regulatory purposes), VINs might be used. However, these scenarios are exceptions rather than the rule and usually involve additional security measures to protect the data.

According to (C2C-CC, 2018), there are a number of measures that have the goal of guaranteeing privacy in V2X messages and their communication:

- **Pseudonymisation**: This is to ensure that the data transmitted cannot be directly linked to a specific individual. This involves using pseudonyms that can only be related to an individual through the collusion of two certification authorities, and only if these authorities have archived the relevant information. This helps in maintaining privacy while still allowing necessary data transmission for safety purposes.
- **Controlled data elements**: The data elements in, e.g., CAMs are carefully selected to exclude any information that can directly identify a vehicle, its owner, or its driver. As mentioned before, data like license plates, registration information, VINs, and other so-called persistent identifiers are not included. This minimises the risk of personal identification from the transmitted messages.
- **Frequently changing identifiers**: In order to prevent tracking and location linking, protocol identifiers are frequently changed during trips. This practice ensures that the continuous reception of V2X messages from the same vehicle does not allow for the reconstruction of a vehicle's journey.
- **Limited data retention**: The received messages are not retained longer than necessary. For example, driving conditions data are kept only for a few seconds to minutes, depending on the service's needs, and are erased once the emission conditions are over.
- **Data minimisation and frequency control**: The frequency of transmission is minimised to the bare essential where possible so as to balance privacy and safety. Typically, European standards reduce the transmission rate to the necessary minimum, considering the driving situation and vehicle speed.
- **Silent periods**: By introducing silent periods between certificate changes, we can mitigate the risk of tracking by making it more difficult to link consecutive messages to the same vehicle and thus reduce the likelihood of continuous tracking.
- **Non-relay of messages**: In order to prevent widespread tracking, received CAMs are not forwarded, nor multi-broadcasted. This restriction limits the data's reach, ensuring that only vehicles within immediate proximity can access the information.

- **Segregation of duties**: The system design incorporates segregation of duties among different authorities to control access to data. The linking of Authorisation Tickets (ATs) to Enrolment Certificates (ECs) is managed by the Authorisation Authority, while linking ECs to vehicle communication unit numbers is handled by the Enrolment Authority. This separation ensures that no single entity can track a vehicle without colluding with multiple authorities, thereby enhancing privacy protection.

In any case, the IEEE and ETSI have established comprehensive standards and protocols for V2X communications, emphasising the protection of privacy through the use of temporary and changing identifiers (as previously mentioned). IEEE standards such as IEEE 1609.2-2016[26] and IEEE 802.11p[27] support secure communication by implementing pseudonym certificates, which are temporary identifiers that periodically change to prevent the tracking of vehicles. These pseudonym certificates authenticate messages without revealing the vehicle's true identity. Similarly, ETSI standards like ETSI EN 302 637-2[28] (CAM) and ETSI EN 302 637-3[29] (DENM) mandate the use of temporary Station IDs that frequently change, ensuring that vehicles cannot be tracked over time. Both sets of standards include provisions for encryption and authentication to enhance the security of V2X communications.

Note that whereas the previously listed headers are primarily metadata that describe the structure, origin, and handling of the message, the payload contains the actual data that the message is conveying, which is often more detailed, sensitive, and specific than the headers[30]. Such data may encompass precise location data (latitude and longitude of vehicles, objects, or events), movement Information (speed, heading, acceleration, and path details), vehicle dimensions: (length and width), event information (details about detected events or environmental conditions), manoeuvre details (intended paths, waypoints, and manoeuvre plans), traffic signal information (current state, phase, and timing of traffic signals), and personal safety information (data related to pedestrians or cyclists, including their positions and movements).

[26] IEEE 1609.2-2016: Standard for Wireless Access in Vehicular Environments (WAVE) – Security Services for Applications and Management Messages.
[27] IEEE 802.11p-2010: IEEE Standard for Information technology – Local and metropolitan area networks – Specific requirements – Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 6: Wireless Access in Vehicular Environments.
[28] ETSI EN 302 637-2 V1.3.1: Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service.
[29] ETSI EN 302 637-3 V1.3.1: Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 3: Specifications of Decentralized Environmental Notification Basic Service.
[30] Note that certain elements like location, speed, and heading appear in both headers and payloads because they are crucial for both the identification and the content of the message. However, the payload often includes more granular and context-specific details, which are necessary for the intended V2X application (e.g., safety warnings, manoeuvre coordination).

4.3 Correlations with personal data

Based on the information from Section 4.2, we now assess the correlation of these V2X-related attributes with possible personal data. Already we can identify the following types of personal information that can be deduced, leaked, or inferred (which are a superset of by (ADV, 2022) earlier identified personal characteristics).

- **Vehicle characteristics**, location and movement information: e.g., latitudes, longitudes, altitudes/elevations, speed, headings, steering angles, path and waypoints, accelerations, brake system statuses, transmission states, vehicle lengths and widths, etc.
- **Vehicle and device identifiers**: these are typically the Station IDs.
- **Temporal information**: these encompass generation times, response times, etc.
- **Intersection and roadway information**: intersection IDs and names, lane lists, etc.

As such, V2X messages can expose various aspects of personal information, primarily concerning location tracking, movement patterns, vehicle and device identification, temporal activities, and interaction with infrastructure.

Based on this analysis, we can highlight what types of correlations there are with personal data. These correlations can be made with different types of personal data, leading to inferences about individuals' behaviours, preferences, and identities.

- **Location and movement patterns, intersection and roadway Information**:
    - **Correlation with home and work locations**: repeated locations in V2X messages (e.g., frequent stops at specific coordinates) can be correlated with an individual's home and workplace.
    - **Travel habits**: patterns of movement (e.g., frequent routes, departure and arrival times) can reveal daily routines, preferred routes, and travel habits.
    - **Real-time tracking**: Continuous location updates can allow real-time tracking of an individual's movements.
    - **Frequent routes**: Intersection data can be used to map out frequent routes, indicating preferred paths and areas of frequent travel.
    - **Travel times**: Data on intersection crossing times can be correlated with travel time analyses, indicating commute durations and preferred travel windows.

- **Device and vehicle identifiers**:
    - **Linking temporary IDs to persistent IDs**: If not frequently-enough changed, then Temporary IDs in BSMs or Station IDs in CAMs can be correlated over time to create a persistent identifier, enabling long-term tracking of a vehicle.
    - **Correlation with vehicle registration**: Identifiers can be cross-referenced with vehicle registration databases to link the vehicle to the owner's personal information (although the likelihood of this is small because the standards prevent the use of this kind of persistent IDs, as explained in Section 4.2.2).
    - **Inferring social relationships**: Identifiers from multiple vehicles frequently seen together can suggest relationships between the vehicle owners.

- **Temporal information**:
  - **Daily schedules**: Generation times and timestamps can be correlated with personal schedules, identifying patterns like work hours, leisure activities, and other time-specific behaviours.
  - **Event participation**: Presence at specific times and locations can indicate participation in events, activities, or social gatherings.

- **Vehicle characteristics and states**:
  - **Driving behaviour**: Data such as speed, acceleration, and braking patterns can be analysed to infer driving behaviour, which can correlate with personal traits like risk tolerance or aggressiveness (note that this is typically exploited in the case of insurance companies providing better value-for-money propositions in case people allow themselves to be monitored).
  - **Vehicle type and preferences**: Information about vehicle size, type, and control settings can be correlated with personal preferences, financial status, and lifestyle choices.

- **Communication metadata**:
  - **Message patterns**: Frequency and patterns of message transmissions can be analysed to determine usage intensity, which can correlate with personal behaviour patterns (e.g., heavy vehicle usage might correlate with certain professions or lifestyles).

4.4 Potential impact on data subjects

Given the concern that increased local processing power and advances in data analytics could make re-identification and privacy issues more pronounced in the future, we now focus on additional potential correlations and privacy risks associated with V2X data that have potential impacts on data subjects, based on the insights obtained in Sections 4.2 and 4.3.

4.4.1 Behavioural biometrics and re-identification risks

A driving style, or even walking trait for pedestrians, can serve as a unique biometric identifier. Distinctive behaviours such as acceleration patterns, braking habits, and steering manoeuvres can be used to distinguish individual drivers. These are not just theoretical; they provide a practical means of identifying drivers even when traditional identifiers are absent. Moreover, vehicle usage patterns, including the frequency, duration, and type of trips taken, can build a comprehensive profile of an individual's lifestyle and habits.

Re-identification risks are significant, especially when location data is involved. By analysing traces, it is possible to re-identify individuals, particularly when this data is combined with other sources like public records or social media check-ins. Despite pseudonymisation, consistent patterns in vehicle or user IDs can reveal the identity of the driver or vehicle owner through sophisticated data linkage techniques.

4.4.2 Cross-referencing with other data sources and behavioural profiling

C-ITS data, when integrated with smart home devices, can provide a detailed picture of an individual's daily schedule. For instance, correlating vehicle data with smart security systems can reveal when a car leaves or arrives at home, thus mapping out daily routines. This integration can extend to social media and public records, where V2X data can be cross-referenced with social media check-ins, photos, or event attendance records. Such cross-referencing can aid in re-identification and provide a richer context for profiling. Behavioural profiling becomes more profound with speed and acceleration patterns. These patterns can indicate aggressive driving, frequent speeding, or adherence to traffic laws, which can lead to detailed profiling of driving habits. Regular travel patterns can also reveal daily routines, personal habits, and lifestyle choices. By understanding these patterns, it is possible to create a comprehensive behavioural profile of an individual.

4.4.3 Temporal and spatial analysis with location-based inferences

Long-term collection and analysis of movement data can create detailed histories of individual travel patterns. Such detailed movement histories enable the prediction of future movements and behaviours, providing insights into an individual's routines and preferences and hence allow us to build a comprehensive picture of where and when an individual travels. Tis type of sensitive information can also encompass home and work addresses, frequent overnight parking locations, daily commute endpoints, and visited locations can even lead to insights on affiliations and personal interests (e.g., regular visits to hospitals or places of worship can indicate health conditions, see also Section 4.4.6, religious beliefs, or other personal interests).

4.4.4 Event participation and inferred social connections

DENMs may uncover proximity to or even involvement in road incidents. Being associated with frequent incidents can mark an individual as a high-risk driver, influencing their insurance premiums and possibly their legal liability in traffic incidents. Furthermore, data showing vehicles frequently in proximity can suggest relationships or interactions between drivers. For instance, vehicles that often travel together can imply that the drivers are family members, friends, or colleagues. Shared routes and regular stops at the same locations can imply carpooling, shared commutes, or common destinations. Such data can hint at social or professional connections, building a network of inferred social interactions based on travel patterns.

4.4.5 Economic and financial inferences with vulnerable road users

The type, age, and maintenance records of a vehicle can provide insights into the owner's economic status and financial health. For instance, owning a luxury vehicle and frequent servicing at high-end dealerships can indicate a higher economic status. Conversely, older vehicles with minimal maintenance can suggest financial constraints. Travel and shopping habits also reveal economic preferences; frequent trips to luxury shopping districts or budget supermarkets can indicate spending habits and economic preferences.

Similarly, PSMs allow to identify patterns and behaviours of vulnerable road users such as pedestrians and cyclists. These patterns raise privacy concerns, especially if the data is used for profiling or targeted advertising. For instance, frequent movement patterns of pedestrians and cyclists can be exploited for commercial purposes or malicious activities; see also Section 4.4.7

4.4.6 Health and wellness indicators from predictive analytics

Regular trips to medical facilities can indicate ongoing medical treatments or chronic health conditions. This type of data is particularly sensitive as it can reveal health issues that individuals might prefer to keep private. Moreover, analysis of travel frequency and destinations such as gyms, parks, or recreational areas can infer activity levels and fitness habits. These inferences can paint a detailed picture of an individual's health and wellness.

4.4.7 Enhanced personal threats

Detailed movement and behavioural profiles can enable intrusively targeted advertising based on frequent routes or destinations, or even predicting when individuals are likely to be in certain locations, increasing the risk of stalking, theft, or even physical harm as criminal activities.

4.5 Summary

Our impact study looked at the potential privacy risks associated with C-ITS data. It primarily investigated the types of information about road users that could be leaked and the potential impacts on individuals. We first examined the specific content transmitted in C-ITS messages, including required and optional headers, message types, and transmission frequencies. The goal was to understand how this information could correlate with personal data, considering future advancements in local processing power and data analytics that could facilitate re-identification.

We then identified various V2X message types used in C-ITS. Each one of them contains specific headers and payloads that include information like vehicle location, speed, heading, and other critical data. These messages are structured to ensure interoperability, with certain common headers. Despite measures like pseudonymisation and frequently changing identifiers, we highlighted potential privacy concerns due to the detailed data transmitted.

In assessing the correlation of V2X attributes with personal data, we furthermore identified several types of personal information that can be deduced from C-ITS messages. These include vehicle characteristics, location and movement information, device and vehicle identifiers, temporal information, and interaction with infrastructure. For instance, repeated location data can reveal home and work locations, travel habits, and real-time tracking capabilities. Device and vehicle identifiers, if not frequently changed, could lead to persistent tracking and inferences about social relationships. Temporal data could uncover daily schedules and event participation, while vehicle characteristics might indicate driving behaviour and personal preferences.

The potential impacts on data subjects are significant, with concerns about re-identification and behavioural profiling becoming more pronounced. Some data can serve as unique biometric identifiers, making re-identification easier when combined with other data sources. Temporal and spatial analyses can reveal detailed travel histories, predict future movements, and infer personal interests and affiliations.

While C-ITS data offers significant benefits for road safety and traffic management, it also poses considerable privacy risks. With advancements in processing power and data analytics, the ability to correlate and analyse C-ITS data will likely improve, making re-identification easier and more accurate. Addressing these concerns requires a proactive approach to data protection and privacy preservation. This underscores the need for evolving stringent data protection measures, anonymisation techniques, and policies to mitigate privacy risks. In conclusion we see that the same principles and mitigation measures, as we identified in Section 3.3, return time and again, and in various fields, e.g., also for tele-operated driving (Bundesanstalt, 2023).

It is noteworthy to mention that impacts could also be qualified by setting up an information classification framework, in which impacts can be categorises into different levels for both organisations and individuals. An example of this is done by (ADVb, 2022), leading to negligible impact, minimal impact, significant impact, serious impact, and threatening impact. However, the qualitative assessment of certain privacy risks in light of re-identifications to corresponding impacts remains a matter of debate and consensus.

5 Current measures and additional measures to reduce risks

tbd

6 Pitfalls that would increase the risk of re-identification

tbd

7 Future recommendations

tbd

8 References

Agentschap Digitaal Vlaanderen, (2022). **Definitie van de standaard data types persoons-gegevens**. Informatieclassificatie Vlaamse overheid (Vo-ICR), Team Informatieveiligheid, 2017-2022.

Agentschap Digitaal Vlaanderen, (2022). **Organisatie informatieclassificatieraamwerk**. Informatieclassificatie Vlaamse overheid (Vo-ICR), Team Informatieveiligheid, 2017-2022.

AFB. (2020). **Mobilidata – Studie naar PKI vereisten**, Versie 2.01, Vlaamse overheid, AWV Mobilidata, Agentschap Facilitair Bedrijf.

Berndt-Tolzmann, S., Burkert, A., Drees, H., Geissler, T., Godarzi, F., and Tarkainen, M. (2022). **Recommendations for continuation of C-ITS deployment**, Deliverable 8 of sub-activity 4.4 (Cooperative ITS Services Deployment Support), EU EIP European ITS Platform.

Boudguiga, A., Stan, O., Fazzat, A., Labiod, H., and Clet, P.-E. (2021). **Privacy Preserving Services for Intelligent Transportation Systems with Homomorphic Encryption**, Proceedings of the 7th International Conference on Information Systems Security and Privacy (ICISSP 2021), pages 684-693.

Bundesanstalt (für Straßenwesen), (2023). **Abschlussbericht der Arbeitsgruppe "Forschungsbedarf Teleoperation"**. Bundesanstalt für Straßenwesen.

Cai, Z., and Xiong, A., (2023). **Understand Users' Privacy Perception and Decision of V2X Communication in Connected Autonomous Vehicles**, Proceedings of the 32nd USENIX Security Symposium, Anaheim, CA, USA.

C2C-CC (2018). **FAQ Regarding Data Protection in C-ITS**, CAR 2 CAR Communication Consortium, 18 September 2018.

Carter, J. M., and Ferber, A. E. (2019). **Using Map Matching for DeIdentification of Connected Vehicle Locations**, IEEE Consumer Electronics Magazine, volume 8, number 6, pages 111-116.

Case, L. (2023). **Advantages and Disadvantages of Homomorphic Encryption**, Baffle.

Chellapandi, V. P., Yuan, L., Żak, S. H., Wang, Z. (2023). **Federated Learning for Connected and Automated Vehicles: A Survey of Existing Approaches and Challenges**, 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, pages 2485-2492.

C-Roads Platform (2021). **Data protection in C-ITS**, Working Group 1 Webinar Results, version 0.95, 19 February 2021.

C-Roads Platform (2023). **C-ITS Message Profiles**, Working Group 2 Technical Aspects, Taskforce 3 Infrastructure Communication, version 2.1.0, 14 December 2023.

C-Roads Platform (2024). **C-ITS Roadmap**, Working Group 2 Technical Aspects, Taskforce 2, version 2.1.0, 18 April 2024.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V.D. (2013). **Unique in the Crowd – The privacy bounds of human mobility**, Scientific Reports.

De Vuyst, C., Smekens, J., Van Aken, K., and Lowet, N. (2022). **Cryptografische maatregelen, Minimale maatregelen, Informatieclassificatie Vlaamse overheid (Vo-ICR)**, Team Informatieveiligheid, Digitaal Vlaanderen.

Dwork, C., McSherry, F., Nissim, K., and Smith., A. (2006a). **Calibrating noise to sensitivity in private data analysis**, Proceedings of the Third conference on Theory of Cryptography (TCC'06), pages 265–284, Springer-Verlag, Berlin, Heidelberg.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006b). **Our data, ourselves: Privacy via distributed noise generation**, Advances in Cryptology, EUROCRYPT 2006, pages 486–503. Springer-Verlag, Berlin, Heidelberg,

EC. (2017). **Opinion 032017 on Processing personal data in the context of Cooperative Intelligent Transport Systems (C-ITS)**, Article 29 Data Protection Working Party.

EC. (2020a). **A European strategy for data**, European Commission.

EC. (2020b). **Ethics of connected and automated vehicles**, Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility (E03659). Ethics of Connected and Automated Vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility. Publication Office of the European Union: Luxembourg.

EC. (2023). **Security Policy for Deployment and Operation of European Cooperative Intelligent Transport Systems (C-ITS)**, Release 3.0, European Commission, Ispra, JRC133795.

EDPB. (2020). **Guidelines on processing personal data in the context of connected vehicles and mobility related applications**, version 2.0, adopted March 2021, European Data Protection Board.

Eland, A. (2015). **Tackling Urban Mobility with Technology**, Google Europe Blog. https://europe.googleblog.com/2015/11/tackling-urban-mobility-with-technology.html

ETSI. (2022). **Intelligent Transport Systems (ITS); Security; Trust and Privacy Management; Release 2**, ETSI TS 102 941 V2.2.1 (2022-11).

Gao, J., Sun, L., and Cai, M. (2019). **Quantifying privacy vulnerability of individual mobility traces: A case study of license plate recognition data**, Transportation Research Part C: Emerging Technologies, volume 104, pages 78-94.

ISO/IEC. (2022). **ISO/IEC 15408-2:2022: Information security, cybersecurity and privacy protection, Evaluation criteria for IT security, Part 2: Security functional components**.

Kamola, M. (2015). **Protecting privacy of GPS trails by anonymization of the road graph**, UrbanGIS@SIGSPATIAL, pages 59-62.

Kapp, A., Hansmeyer, J., and Mihaljevic, H. (2023a). **Generative Models for Synthetic Urban Mobility Data: A Systematic Literature Review**, ACM Computing Surveys, volume 56, number 4, Article 93.

Kapp, A., and Mihaljevic, H. (2023b). **Reconsidering utility: unveiling the limitations of synthetic mobility data generation algorithms in real-life scenarios**, SIGSPATIAL '23, Hamburg, Germany.

Lian, J., Wang, D., Zhu, S., Wu, Y., and Li, C. (2022). **Transformer-Based Attention Network for Vehicle Re-Identification**, Electronics, volume 11, number 7, pages 1007-1016.

Liu, X., Liu, W., Mei, T., and Ma, H. (2018). **PROVID: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance**, IEEE Transactions, Multimedia volume 20, number 3, pages 645-658.

Malin, B., and Sweeney, L. (2004). **How (not) to protect genomic data privacy in a distributed network: Using trail re-identification to evaluate and design anonymity protection systems.** Journal of Biomedical Informatics, volume 37, number 3, pages 179–192.

Maouche, M. (2019). **Protection against re-identification attacks in location privacy**, Networking and Internet Architecture, Université de Lyon.

Mehner, L., von Voigt, S. N., and Tschorsch, F. (2021). **Towards Explaining Epsilon: A Worst-Case Study of Differential Privacy Risks**, 2021 IEEE European Symposium on Security and Privacy Workshops (EuroS & PW).

Nelson, G. (2015). **Practical Implications of Sharing Data - A Primer on Data Privacy, Anonymization, and De-Identification**, ThotWave Technologies, Chapel Hill, NC, paper 1884-2015.

Pyrgelis, A., Kourtellis, N., Leontiadis, I., Serrà J., and Soriente, C. (2018). **There goes Wally: Anonymously sharing your location gives you away**, 2018 IEEE International Conference on Big Data, Seattle, WA, USA, 2018, pages 1218-1227.

Raes, L., Stott, A., and Versmissen, R. (2020). **Privacy rules and data anonymisation**, Deliverable 4.7, PoliVisu project, Grant Agreement number 769608.

Rebiger, S., Moraes, T., Lareo López de Vergara, X., and Zerdick, T. (2019). **Connected Cars**, EDPS Tech Dispatch, issue 3.

Rondinone, M. and Correa, A. (2018). **Definition of V2X Message Sets**. TransAID Horizon 2020 Deliverable D5.1.

Tan, Z., Wang, C., Fu, X., Cui, J., Jiang, C., and Han, W. (2017). **Re-identification of Vehicular Location-Based Metadata**, ICST Transactions on Security and Safety.

Tan, C., and Yang, K. (2024). **Privacy-Preserving Adaptive Traffic Signal Control in a Connected Vehicle Environment**, Transportation Research Part C: Emerging Technologies, volume 158, 104453.

Tao, Y., Javanmardi, E., Lin, P., Nakazato, J., Jiang, Y., Tsukada, M., and Esaki, H. (2023). **Zero-Knowledge Proof of Traffic: A Deterministic and Privacy-Preserving Cross Verification Mechanism for Cooperative Perception Data**. IEEE Access, pages 1-1. 10.1109.

Zakria, Deng, J., Z., Hao, H., Khokhar, M.S., Kumar, R., Cai, J., Kumar, J., and Aftab, M.U. (2021). **Trends in Vehicle Re-Identification Past, Present, and Future: A Comprehensive Review**, Mathematics 9, number 24: 3162.

Zhao, C., Zhao, S., Zhao, M., Chen, Z., Gao, C.-Z., Li, H., and Tan., Y. (2019). **Secure Multi-Party Computation: Theory, Practice and Applications**, Information Sciences, volume 476, pages 357–372.

Zheleva, E., and Getoor, L. (2007). **Preserving the privacy of sensitive relationships in graph data**. In Proceedings of the 1st ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD (PinKDD '07). ACM.

Appendix A Insights from experts

A.1 Expert workshop #2 (14/05/2024)

A.1.1 Background and context

On 14 May 2024, we held an online workshop with several invited experts, to whom we presented the following relevant questions:

- **General**
  - What are the **primary risks** associated with vehicle re-identification in the context of connected vehicles, and how might these risks evolve in the next five years? Are we fighting an **uphill battle**?
  - What are the most challenging **technical hurdles** currently facing developers trying to secure connected vehicles against unauthorised data access and re-identification?

- **Regulations and ecosystems**
  - How do **current regulatory frameworks** address the issues of vehicle re-identification and connected vehicle privacy? Are there **specific gaps** that need to be filled?
  - How can stakeholders in the connected vehicle ecosystem collaborate to enhance vehicle data security and user privacy? What **models of collaboration** have been successful in other industries?
  - What are the potential **impacts** of vehicle re-identification **on insurance industries and law enforcement**? How should these sectors prepare for these impacts?

- **Practical and ethical**
  - In terms of connected vehicle data, what is the **balance between utility and privacy**? How should this balance be managed or regulated?
  - What **ethical considerations should guide** the development and implementation of technologies aimed at protecting against vehicle re-identification and deanonymisation?

As such, the workshop on WP4 focused on various aspects of privacy and security in connected vehicle ecosystems. The discussions covered primary risks and challenges associated with vehicle re-identification, technical hurdles in securing connected vehicles, and the implications of opting out of C-ITS services. Additionally, the workshop examined the regulatory frameworks, collaborations within the ecosystem, and the ethical considerations in developing and deploying these technologies.

In order to get the group going, we organised a Miro board that allowed attendants to easily provide their own opinions and feedback on each specific question.

A.1.2 Miro boards

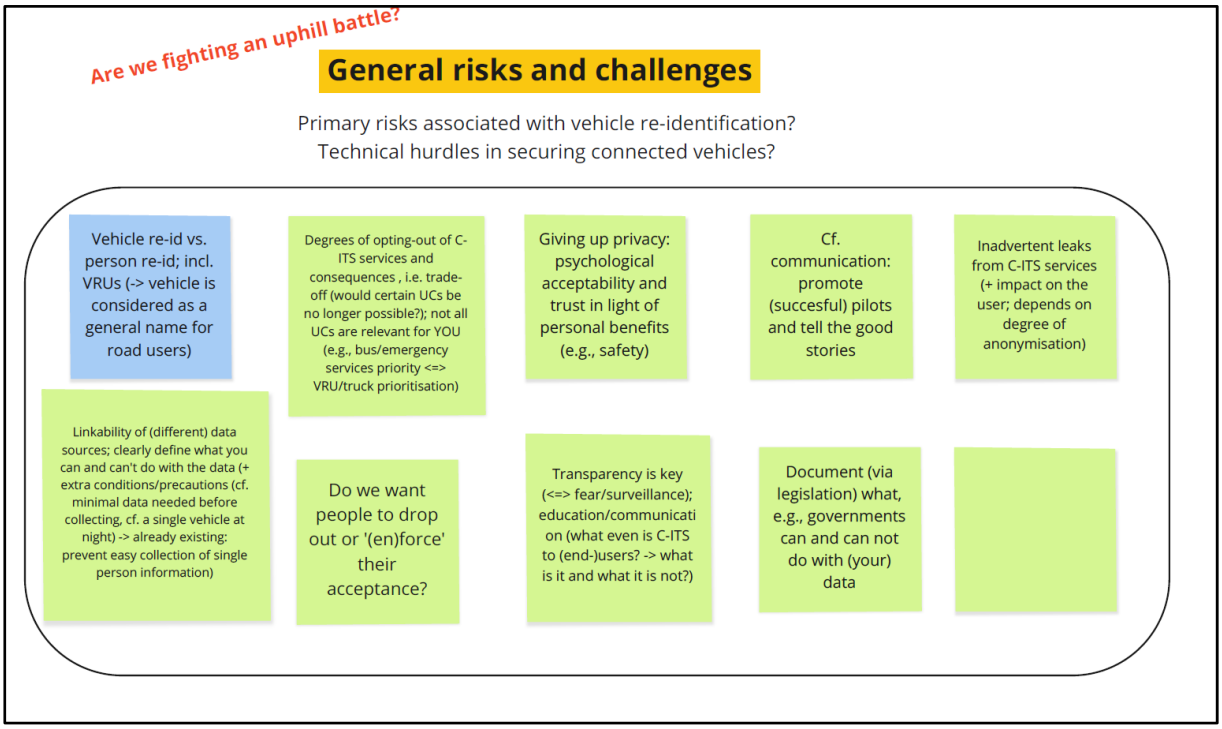

Figure 1: Expert workshop Miro board for topics related to general risks and challenges.

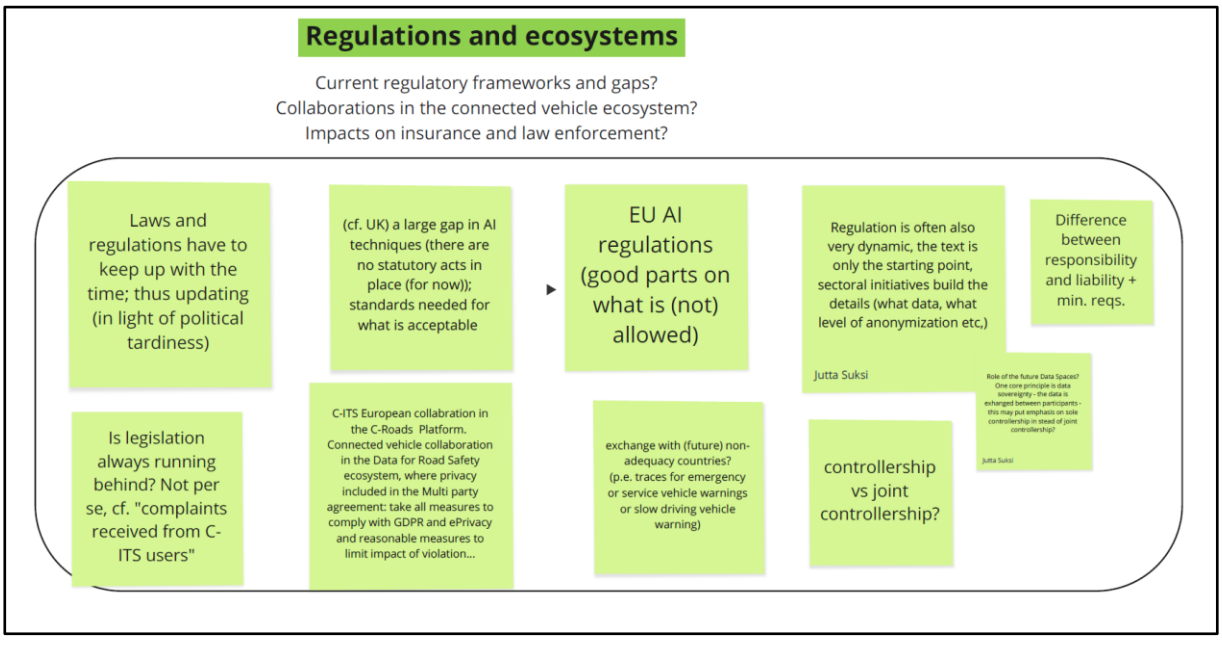

Figure 2: Expert workshop Miro board for topics related to regulations and ecosystems.

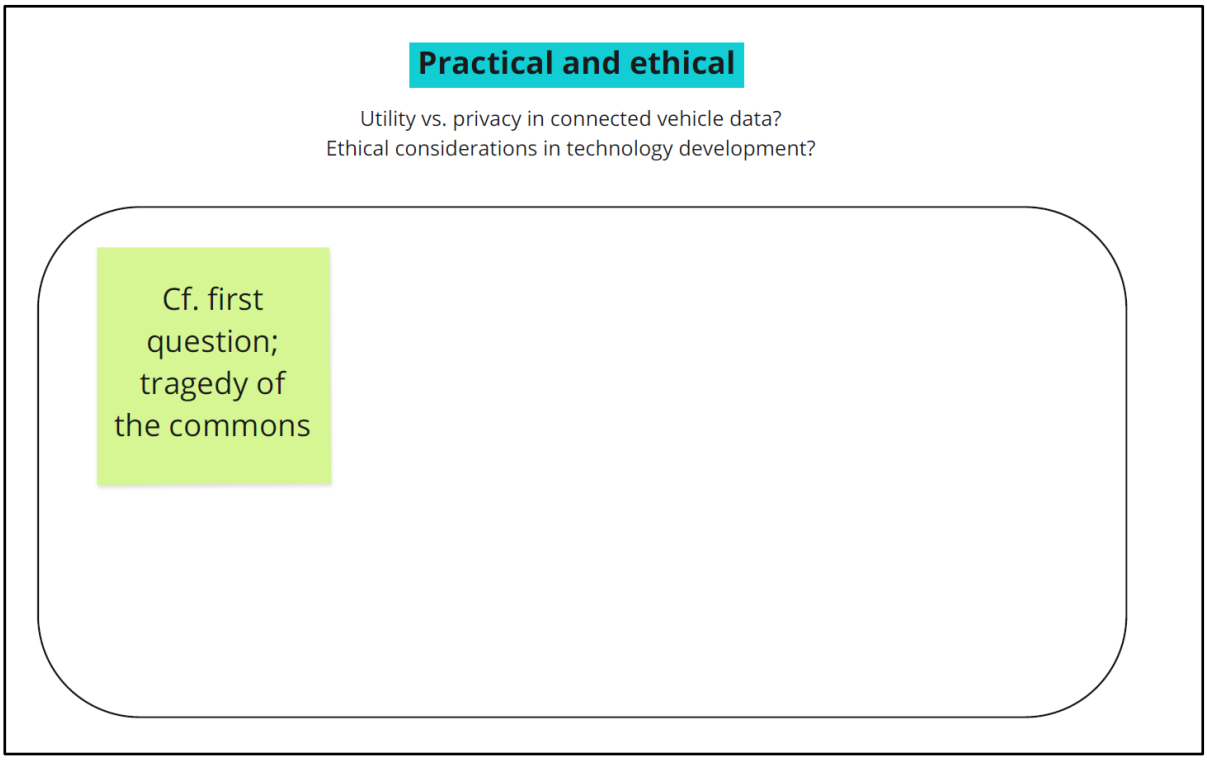

Figure 3: Expert workshop Miro board for topics related to practical and ethical issues.

Note that due to time constraints, the third discussion topic (Practical and ethical) was not fully covered.

A.1.3 Summaries from the workshop

During our workshop, we dived in three main topics: general risks and challenges, regulatory frameworks and gaps, and practical and ethical considerations. The discussion was enriched with detailed insights, diverse perspectives, and thorough observations, reflecting the intricate and multifaceted nature of data privacy and even AI-related issues.

A.1.3.1 General risks and challenges

We began by delving into the risks associated with vehicle re-identification. This concern primarily revolves around the ability to trace specific vehicles and, by extension, their owners through data collected by various services. Participants pointed out that techniques for vehicle identification are sophisticated, involving not just number plates but also patterns in driving behaviour and vehicle usage. One attendee highlighted the risk of ‘data mosaicking’, where disparate data points are combined to create a comprehensive profile of an individual.

Moreover, the group discussed broader implications of data collection, such as the potential misuse of information by both governmental and private entities. There was consensus that the aggregation of data from multiple sources exacerbates these risks, as it increases the chances of re-identification and misuse. A participant raised the point that in regions with less stringent data protection laws, the risks are even more pronounced, making international data transfers a significant concern.

A furthermore critical challenge discussed related to the data lifecycle itself, from collection to storage to eventual deletion, which presents numerous opportunities for breaches and misuse. Participants noted that each stage of this lifecycle must be managed with stringent security measures to prevent unauthorised access and potential exploitation. There was also a discussion about the increasing capability of AI to analyse vast datasets, making it easier to identify patterns and infer sensitive information.

The psychological aspect of privacy was another nuanced point. Attendees observed that people are often willing to trade some degree of privacy for tangible benefits, such as improved safety and convenience. This led to a discussion about the importance of making these benefits explicit to the public to garner their consent and cooperation. However, the trade-off must be transparent and justifiable, ensuring that individuals are fully aware of what they are giving up and what they stand to gain.

Inadvertent data leaks from C-ITS services were identified as a significant risk, particularly when anonymisation is insufficient. The degree of impact on users depends on how well the data is anonymised. Participants advocated for legislative measures to clearly document the permissible actions of governments and other entities with user data. This legal clarity can help mitigate the risk of unauthorised data collection and usage.

A.1.3.2 Regulatory frameworks and gaps

The discussion on regulatory frameworks revealed significant gaps and challenges. The European AI Act was mentioned as a progressive step, but it was acknowledged that regulatory measures often lag behind technological advancements. One speaker emphasised that regulation is inherently reactive, catching up with innovations rather than anticipating them. This lag can result in temporary regulatory vacuums where new technologies operate without sufficient oversight.

The attendees discussed the need for a more agile and dynamic approach to regulation. This could involve periodic reviews and updates to legislation to ensure it remains relevant and effective. The idea of regulatory sandboxes was floated, allowing for controlled experimentation with new technologies under regulatory supervision. This could help in identifying potential issues early and refining regulations accordingly.

International regulatory disparities were another critical point. The differences between the GDPR in Europe and the evolving data protection laws in the UK post-Brexit were discussed in detail. One participant pointed out that these differences create compliance challenges for companies operating across borders. There was a call for greater harmonisation of data protection laws (globally) to ensure consistent standards and reduce the complexity of compliance. The European Union's AI regulations were praised for their comprehensive approach to defining permissible actions, though gaps remain, especially in regions like the UK, where statutory acts on AI techniques are still lacking.

Collaboration within the connected vehicle ecosystem is crucial for effective regulation. The C-Roads Platform and the Data for Road Safety ecosystem were cited as examples of successful initiatives where privacy is integrated into multi-party agreements. These collaborations ensure compliance with GDPR and ePrivacy regulations, emphasising the need for continual adaptation and sectoral initiatives to detail data usage protocols.

A particularly interesting observation was made regarding the role of AI in regulatory compliance itself. Some attendees suggested that AI could be used to monitor and enforce compliance with data protection regulations, providing a proactive approach to identifying and addressing breaches. This use of AI could enhance the effectiveness of regulatory frameworks by automating the detection of non-compliance and alerting authorities in real-time.

The workshop also delved into the roles of controllership and joint controllership in data management. With the increasing emphasis on data sovereignty, the distinction between sole and joint controllership becomes critical. The future role of Data Spaces, where data is exchanged between participants, could shift the balance towards sole controllership, highlighting the need for clear regulatory definitions and responsibilities.

A.1.3.3 Practical and ethical considerations

The final segment of the workshop focused on practical and ethical considerations. Transparency emerged as a central theme, with broad agreement that organisations must be clear about what data they collect, how it is used, and the benefits it provides. This transparency is crucial for building and maintaining public trust. An attendee highlighted the importance of not just legal compliance but ethical behaviour, suggesting that companies should go beyond the minimum requirements of the law to protect user privacy.

User consent and the ability to opt-out were extensively debated. While consent is a cornerstone of data protection laws, the ease and clarity with which users can give or withdraw consent are often lacking. One participant shared an example of a mobile app that made it very difficult for users to opt-out of data collection, illustrating a gap between theoretical consent and practical implementation. There was consensus that consent mechanisms need to be user-friendly and transparent, enabling individuals to make informed choices about their data.

The role of governments and private companies in ensuring ethical data practices was another key point. Attendees noted that governments have a responsibility to enforce regulations and provide oversight, but private companies must also take proactive steps to protect data. This includes implementing robust security measures, conducting regular audits, and being transparent about data breaches when they occur. Hence, there should also be a focus on defining clear roles and responsibilities.

Balancing utility and privacy in connected vehicle data is a complex challenge. The utility of data for improving traffic management, enhancing safety, and enabling new services must be weighed against the potential invasion of privacy. Ethical considerations in technology development were a recurring theme, with discussions on the tragedy of the commons in data usage. This concept illustrates the dilemma where individual actions to maximise personal benefit can lead to collective detriment.

A.1.3.4 Observations and reflections

Throughout the workshop, several observations and reflections were made by attendees. One recurring theme was the need for continuous education and awareness-raising among the public about data privacy issues. This includes not just understanding the risks but also knowing how to protect oneself and make informed decisions about data sharing. Participants noted that public awareness campaigns and educational initiatives could play a crucial role in enhancing data literacy.

Another important point was the need for greater collaboration between different sectors – government, industry, academia, and civil society – to address the challenges of data privacy and AI. This collaboration can lead to more holistic solutions that take into account the diverse perspectives and expertise of various stakeholders. One attendee suggested the formation of multi-stakeholder committees to regularly review and update privacy regulations and guidelines, ensuring they remain relevant in the face of rapid technological change.

There was also discussion about the role of accountability in data privacy. Attendees emphasised the need for clear accountability mechanisms to ensure that organisations adhere to data protection principles and regulations. This includes not only legal accountability but also moral and ethical responsibility. One participant mentioned the concept of privacy by design, where privacy considerations are integrated into the design and development of systems and technologies from the outset.

The issue of data ownership was another significant topic. Who owns the data collected by various devices and services? There was a consensus that individuals should have ownership and control over their data, with the right to access, correct, and delete their information. However, the practical implementation of this principle presents challenges, especially in complex data ecosystems involving multiple actors.

In conclusion, the workshop highlighted the importance of balancing technological innovation with robust regulatory frameworks and ethical considerations, ensuring that the benefits of these technologies are realised without compromising individual privacy and rights. The discussions underscored the need for continuous dialogue, collaboration, and innovation.

A.2 PEB workshop (18/06/2024)

A.2.1 Background

During the PEB meeting the TIARA team held a workshop that centred around certain questions, ranging from topics such as stakeholder identification and literacy, dissemination channels and means, data quality, data privacy, ubiquitous coverage, and PKI. In the following paragraphs, we share the results of the data privacy group.

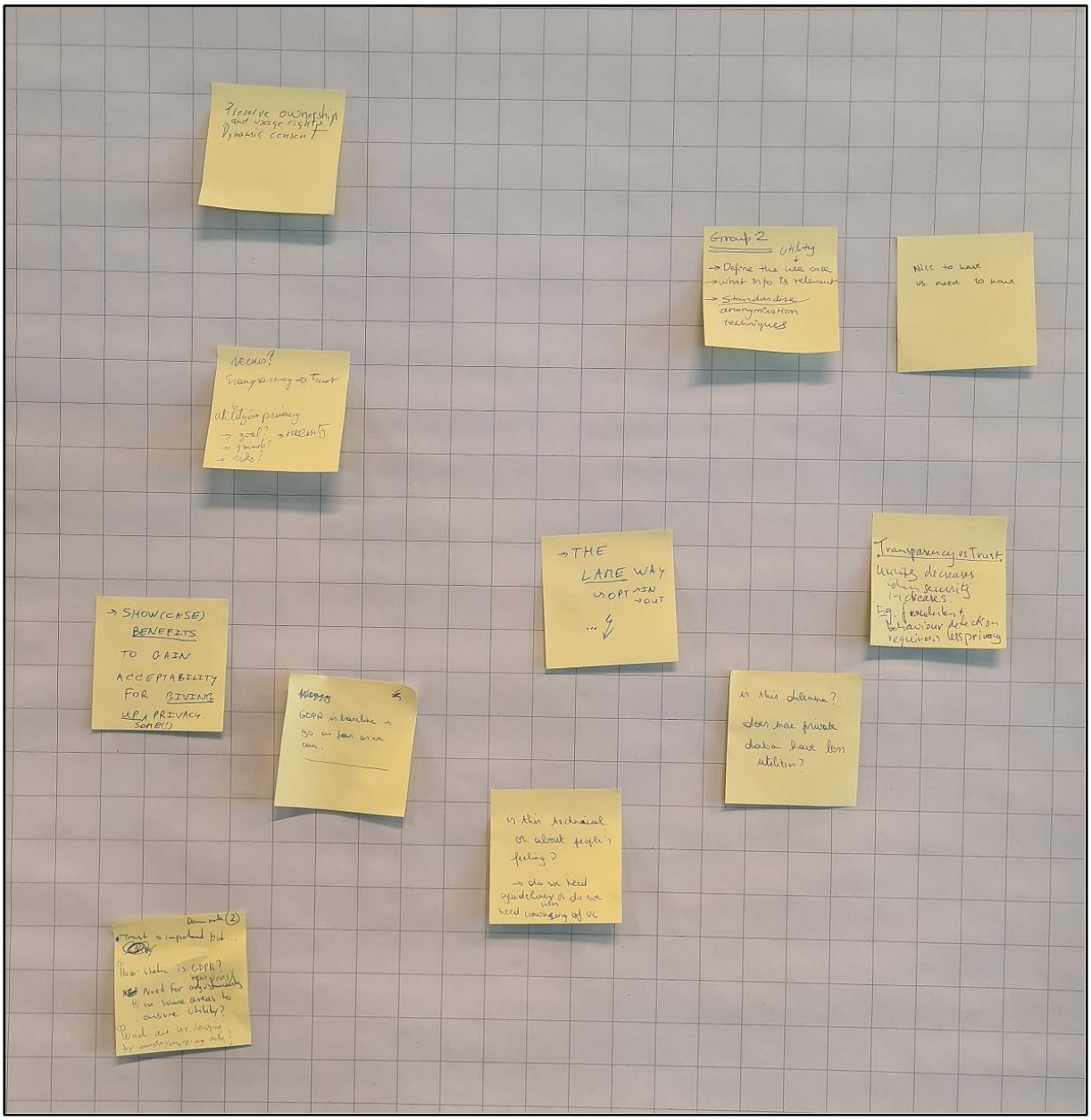

Figure 4: PEB workshop notes board for topics related to privacy concerns.

A.2.2 Summary from the workshop

During the discussions the participants underscored several essential concepts. A key point of emphasis was preserving ownership and usage rights, which is integral to maintaining user trust and ensuring compliance with legal frameworks like GDPR. Dynamic consent was highlighted as a crucial mechanism, enabling users to continuously control how their data is utilised. Combined with transparency, this builds trust and creates a robust foundation for ethical data practices.

A significant portion of the discussion focused on the balance between data utility and privacy. It was deemed imperative to clearly define the goals, grounds, and risks associated with data usage. For data to be useful, the specific use case must be identified, determining which information is relevant and necessary. There was debate over the notion that data utility decreases as security increases, with some arguing that enhanced security measures can, e.g., hinder the detection of fraudulent behaviour due to increased privacy constraints. However, this raises the question of whether more private data inherently possesses less utility, challenging us to explore ways to maximise data utility while maintaining stringent privacy standards.

Finally, the group considered the technical versus emotional aspects of data privacy. While technical guidelines and rules are necessary to establish a baseline, as evidenced by GDPR, the importance of convincing people about the need for privacy adjustments for specific use cases was noted. Standardising anonymisation techniques was suggested as a way to enhance both privacy and utility. Showcasing the benefits of data usage can help gain public acceptability, encouraging individuals to consent to sharing their data. The static nature of GDPR was questioned, with suggestions for adjustments to ensure it keeps pace with evolving data utility needs. Lastly, the implications of data anonymisation were discussed, weighing the potential loss of valuable insights against the necessity of protecting individual privacy.

A.2.3 Comparison with earlier results

The results from the PEB workshop aligned quite well with the information already mentioned in Section 0 as well as the results from the previous expert workshop. In addition, the PEB workshop provided extra insights into preserving ownership and usage rights, dynamic consent, the emotional aspects of data privacy, standardising anonymisation techniques, showcasing the benefits of data usage, questioning the static nature of GDPR, and the implications of data anonymisation.