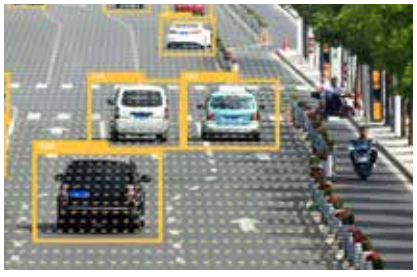
International Transport Forum

CPB
Corporate Partnership Board

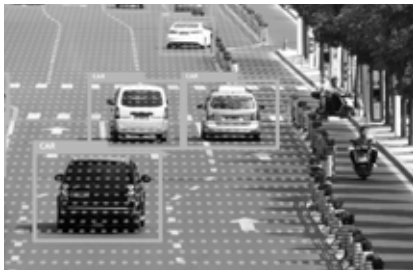
Safer Roads with Automated Vehicles?



Corporate Partnership Board Report

OECD

Safer Roads with Automated Vehicles?



Corporate Partnership Board Report

About the International Transport Forum

The International Transport Forum at the OECD is an intergovernmental organisation with 59 member countries. It acts as a think tank for transport policy and organises the Annual Summit of transport ministers. ITF is the only global body that covers all transport modes. It is administratively integrated with the OECD, yet politically autonomous.

ITF works for transport policies that improve peoples' lives. Our mission is to foster a deeper understanding of the role of transport in economic growth, environmental sustainability and social inclusion and to raise the public profile of transport policy.

ITF organises global dialogue for better transport. We act as a platform for discussion and pre-negotiation of policy issues across all transport modes. We analyse trends, share knowledge and promote exchange among transport decision makers and civil society. ITF's Annual Summit is the world's largest gathering of transport ministers and the leading global platform for dialogue on transport policy.

Our member countries are: Albania, Argentina, Armenia, Australia, Austria, Azerbaijan, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Canada, Chile, China (People's Republic of), Croatia, Czech Republic, Denmark, Estonia, Finland, France, Former Yugoslav Republic of Macedonia, Georgia, Germany, Greece, Hungary, Iceland, India, Ireland, Israel, Italy, Japan, Kazakhstan, Korea, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Mexico, Republic of Moldova, Montenegro, Morocco, Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Russian Federation, Serbia, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine, the United Arab Emirates, the United Kingdom and the United States.

Disclaimer
Funding for this work has been provided by the ITF Corporate Partnership Board. This report is published under the responsibility of the Secretary-General of the ITF. It has not been subject to the scrutiny of ITF or OECD member countries, and does not necessarily reflect their official views or those of the members of the Corporate Partnership Board.

Acknowledgements

The work for this report was carried out in the context of a project initiated and funded by the International Transport Forum's Corporate Partnership Board (CPB). CPB projects are designed to enrich policy discussion with a business perspective. They are launched in areas where CPB member companies identify an emerging issue in transport policy or an innovation challenge to the transport system. Led by the ITF, work is carried out in a collaborative fashion in working groups consisting of CPB member companies, external experts and ITF staff. Many thanks to the members of the Corporate Partnership Board companies involved in this work: Abertis, Alstom, Anheuser-Busch InBev, Brisa, Ford, Here, Inrix, Kapsch TrafficCom, NXP, PTV Group, RATP Group, Renault-Nissan Alliance, Siemens, SNCF (Keolis), Toyota, Transdev, Uber, Volvo Car Corporation, Volvo Group.

The report draws conclusions from the Workshop "Safety and Security on the Road to Automated Transport: The Good, the Uncertain and the Necessary" held in November 2017 in Paris, France. Participants of the workshop included:

Mr. Gerhard ALTMANN, SAS
Dr. Jean-Pascal ASSAILLY, IFSTTAR
Mr. Jean-François BOULINEAU, RATP Group
Prof. Oliver CARSTEN, Institute for Transport Studies, University of Leeds
Mr. Mihai CHIRCA, Transdev
Mr. Philippe FENAIN, Abertis
Mr. Guy FREMONT, Abertis
Mr. Philip KING, Volvo Car Corporation
Mr. Peter KRONBERG, Volvo Group
Mr. Gérard LE LANN, INRIA
Prof. Anders LIE, Trafikverket
Mr. Marius MACKU, Uber
Dr. Gereon MEYER, VDI/VDE Innovation + Technik GmbH
Mr. Luca PASCOTTO, FIA - Federation Internationale de l'Automobile
Mr. Karl PIHL, Volvo Group
Ms. Sofia SALEK DE BRAUN, PTV Group
Mr. Cristian SANDU, AutoKAB
Mr. Markus SOPPA, accessec GmbH
Mr. Kusajima TAKAYUKI, Toyota Motor Corporation
Mr. Jasja TIJINK, Kapsch TrafficCom AG
Mr. Adrian ULISSE, INRIX
Mr. Philippe VENTEJOL, RATP Group

We would like to thank Oliver Carsten (ITS- Leeds), Anders Lie (Trafikverket), Henrik Kiertzner (SAS), Peter Kronberg (Volvo Group), Gerard Le Lann (INRIA), Eva Molnar (formerly UNECE), Gereon Meyer (VDI/VDE Innovation + Technik GmbH), Sebastian Rohr (accessec GmbH), Markus Soppa (accessec), Jasja Tijink (Kapsch TraffiCom), and Philippe Ventejol (RATP)for their contributions to this report.

The principal authors of this report are Philippe Crist and Tom Voege. The project was managed by Tom Voege. Sharon Masterson coordinated CPB activities.

Table of contents

Executive summary

What we did

This report examines how increasing automation of cars and trucks could affect road safety and which security vulnerabilities will need to be addressed with the rise of self-driving vehicles. The report applies the principles of the "Safe System", which is at the forefront of current thinking about road safety, to the wider discussion on vehicle automation. It also takes into considerations the security of the cyber-physical system associated with automated driving. This includes a definition of relevant system boundaries and future-proof minimum requirements for relevant safety and security indicators.

A review of the key issues and discussions in the context of (road) safety and (cyber) security of (highly) automated vehicles framed the discussion at a workshop held in November 2017, which, combined with research and expert input, provides the basis for this report.

What we found

Widespread vehicle automation has never seemed as close as today. The attraction of automated vehicles lies in the compelling promise the technology holds for safer performance, more efficient traffic and the development of new markets. While the potential seems great, there are many unknowns that public authorities must manage. There are also certain pre-requisites that authorities must guarantee.

The first among these pre-requisites is that vehicles and the traffic system must be safe. The principal tenet of the Safe System approach is that traffic systems should be designed in such a way that human fallibility does not result in death or serious injury. Conceived to ensure safety in a world full of human error, the Safe System can also deliver safety in a world of machine errors or unanticipated behaviours.

Claims of a more than 90% reduction in road traffic deaths resulting from automation eliminating crashes linked to human error are untested. It seems likely that the number of road casualties will decrease with automation, but crashes will not disappear. In certain circumstances, more crashes may occur among "average" drivers that are not prone to risky behaviour. This is particularly likely in circumstances where drivers must take over from automated driving in emergency situations.

The lack of experience and data complicates an assessment of how safe automated driving really is. It is further complicated by the lack of a common framework for such a safety performance assessment and by rapid changes in its object: a self-driving car is, after all, a combined hardware and software system whose critical performance characteristics can change radically with software upgrade.

Vehicle automation strategies that keep humans involved in the driving task seem risky. A shared responsibility for driving among both automated systems and humans may not render decision making simpler, but more complex. Thus, the risk of unintended consequences that would make driving less safe, not more, could increase.

Humans retain an advantage over single sensor-based automated systems in many contexts. Overcoming this gap requires combining input from several sensors. In some cases, safe operation will require vehicles to communicate with each other and with infrastructure beyond line of sight. However, to rely on this connectivity for safety performance is fraught with risks, especially with regard to cybersecurity. Whether vehicle automation should move from a reactive safety paradigm (where vehicles rely on their own

capabilities) to a proactive safety framework (where vehicles are embedded in a communicative network) is still debated.

Two fundamental design strategies condition cybersecurity for automated driving. The first relates to the functional isolation of a safety-critical subsystem, the second to whether safe performance is conditioned on connectivity to external networks. These are not trivial design decisions. The choice of strategy will have an incidence on whether imperatives for safety and cybersecurity can be reconciled – and if so, how easily or not.

What we recommend

Reinforce the Safe System approach to ensure automated vehicles are used safely.

The Safe System, or systematic safety, approach to managing the road transport system is holistic and proactive. Vehicles, infrastructure and traffic management combine to guide users to act safely to prevent crashes and when they occur ensure that impact forces do not exceed the physical limits of the body. Automation places even greater importance in achieving an effectively integrated Safe System approach to operation.

Apply Vision Zero thinking to automated driving.

Adopting an outcome-based or performance-based approach as opposed to a prescriptive technology-based approach to safety is in line with many other domains of governance (education, health, etc.). If this is the strategy employed, countries will have to consider what outcome they want the system to deliver – a reduction in deaths and injuries or zero road deaths and serious injuries.

Avoid safety performance being used to market competing automated vehicles.

Regulators and industry should work together to ensure uniform safety performance of automated driving systems. The relative safety level of vehicles deployed, or strategies employed, should not be a competition issue. The regulatory framework should ensure maximum achievable road safety, guaranteed by industry, as a precondition of allowing these vehicles and services to operate.

Carefully assess the safety impacts of systems that share driving tasks between humans and machines.

Automated vehicles designed to share driving tasks between humans and machines have so far confounded efforts to ensure safety. It is unclear to what extent workarounds in system design can effectively address poor task allocation, de-skilling or cognition and control issues. This is partially due to the lack of statistically valid data on the efficacy of these measures, but it is possible that these are intractable vulnerabilities than cannot be "designed away". In that case, the Safe System approach suggests either avoiding machine-to-human handovers or allowing them only when safety performance goals can be demonstrably met i.e. when poorly managed handovers will not result in death or serious injury.

Require reporting of safety-relevant data from automated vehicles.

Nearly all highly automated driving systems currently in testing rely on humans as back-up drivers who will take over when the computer reaches the limits of its performance. The disengagement of automated driving systems, not just crashes, are a relevant metric for tracking safety performance because the vehicle might have crashed without human intervention. Since different automated driving systems have different design parameters and functional boundaries, metadata on system capabilities should accompany

disengagement reports. If the ultimate goal is to understand to what extent automated driving systems can copy, or improve on, human driving skills, the appropriate metric for comparison is the number of near-misses avoided compared to human drivers. Emerging new sources of data may enable this to be measured. The number of kilometres driven in automated mode is more relevant than the number of kilometres driven overall by these vehicles.

Develop and use a staged testing regime for automated vehicles.

In all stages of the development process of self-driving cars from early prototype to market-ready, a robust safety testing regime needs to be in place and strictly adhered to such that developers can demonstrate how, in what contexts and why a solution is safe. Such a staged approach would move from simulation via tests on private tracks to piloting on public roads with a human back-up driver. Only then would the automated vehicle be allowed on public roads. This sequence avoids involving the public in "beta testing", which is incompatible with Safe System thinking.

Establish comprehensive cybersecurity principles for automated driving.

Robust cybersecurity in complex "systems of systems" like automated driving requires comprehensive and shared frameworks. Good examples are the Key Principles of Cyber Security for Connected and Automated Vehicles developed by the Department for Transport (DfT) in the United Kingdom, which outline the fundamental building blocks that should underpin systemic cybersecurity best practice and the Society of Automotive Engineers Cybersecurity Guidebook for Cyber-Physical Vehicle Systems (SAE,2016).

Ensure the functional isolation of safety-critical systems and that connectivity does not compromise cybersecurity or safety.

Connectivity can help improve the situational awareness of automated vehicles and provide input that enhances their safety performance. At the same time, cybersecurity risks for automated driving are similar to other complex systems like airplanes, trains and metro systems. In all of these, core safety-critical components are functionally isolated on both a hardware and software level from non-critical components. Where these functional boundaries lie must be based on robust risk assessment. A second fundamental design principle is that the avoidance of crashes should never depend on access to shared external communication channels alone.

Provide clear and targeted messaging of vehicle capabilities.

The discourse on automated vehicles currently relies on the five levels of automation defined by the Society of Automotive Engineers (SAE) in 2014 to describe technology capabilities. The SAE automation levels are useful and necessary in a technical context. Simpler ways of messaging about what automated cars can and cannot do will have to be developed in order to explain vehicle automation to non-technical audiences. Policy-makers will need to become more aware of what different forms of automation imply for specific policy objectives, such as safer roads. End-users will need help to form realistic expectations about vehicle performance and better understand their role in automated driving.

Introduction

Widespread vehicle automation has never seemed as close as it seems now given the confluence of advanced sensing devices, on-board and remote processing capabilities, geospatial location technologies, ubiquitous connectivity and emerging shifts in the way the cars and other vehicles are used and owned. It is a topic that has garnered a significant level of investment by industry, mobilised broad research interest and is proving to be challenging for regulatory frameworks around the world. The attraction of automated vehicles to these and other stakeholders lies in the compelling promise the technology holds for safer performance, more efficient traffic and new market development.

The far-reaching behavioural and policy implications of vehicle automation both excite and worry regulators and the public, but it is the rapidity with which technological change is occurring in this area that is pushing existing frameworks. The rapid up-scaling of automated driving technologies and capabilities presents a break with past patterns of vehicle technology development that has largely been incremental and focused on relatively well-understood technology platforms. Except at the outset of motorisation in the early 20th century, vehicle and traffic regulations have not been significantly disrupted by new technology concepts and use cases. This is changing with the impending arrival of highly, and eventually fully, automated cars, trucks, buses and other road vehicles.

Advanced trials indicate that some automated cars are already able to operate reliably in certain contexts, but variable performance across the industry suggests that the case for widespread deployment for all vehicles in all contexts and for all uses has yet to be made. Further, there seems to be little consensus on what automation pathways look like vis-à-vis the role of the human driver (controller, supervisor or completely out of the loop). Nor is there consensus on what role automated driving will play in the traffic system of tomorrow (VDA, 2015). Today cars are largely independent technology platforms that are individually owned and operated. The degree with which this remains the case in the future with the advent of full automation is a subject of broad speculation and study. The outcome of that speculation will depend not only on what happens within the industry but will also depend on trends external to the car industry and to the transport sector itself.

Depending on the deployment model for highly automated vehicles, e.g. fleet-based or individually owned, wider impacts of automation may radically reshape transport demand or change the nature of existing demand. Though the potential is great, there are many unknowns that public authorities must manage and certain pre-requisites that authorities must guarantee. First among these is that vehicles must be safe. A second, but equally important consideration is that the traffic system into which automated vehicles are inserted must, at a minimum, not be rendered less safe than it is today. Ideally, of course, system-wide safety should be improved with the uptake of highly automated and fully automated vehicles.

These tasks – assessing vehicle safety and assessing traffic system safety – are at the heart of the regulatory function in the field of transport. They are not new tasks and the imperative to reduce traffic deaths and injuries has given rise to sound assessment frameworks adopted by public authorities. Among these is the "Safe System" approach or systematic safety approach to traffic safety. This report explores if this tested approach to improving road traffic safety can be adapted to automated driving, where this may be challenging, and where vehicle automation can give rise to new concepts for traffic system safety. It does so by:

- Exploring the "Safe System" approach.
- Discussing the potential safety benefits/disbenefits of automated driving.
- Raising the safety considerations related to the cyber-security of vehicles and traffic systems.

The safe traffic system

Road traffic deaths and serious injuries are a confounding and unacceptable outcome of transport activity. The implementation of evidence-based policies has helped to reduce road traffic deaths and injuries around the world (ITF 2008) (ITF 2016) but road traffic still kills and maims thousands of people every year. The World Health Organisation (WHO) estimates that in 2013, 1.25 million people died worldwide as an effect of road accidents and up to 50 million people yearly are injured in road crashes (WHO 2015).

Fatality and injury rates in rapidly developing countries are high compared to rates in countries that motorised earlier and that benefit from more mature transport systems and above all have adopted evidence-based road safety policies. Countries with relatively low traffic fatality rates today have not always displayed good traffic safety performance. In the 1960s and 1970s some of the best performing countries today had fatality rates in line with some of the worst today (TCS, 2008).

The experience of countries achieving the largest drops in fatality and serious injury rates has highlighted what policies are most effective. A common feature of several well-performing countries – despite differences in terminology and operationalisation – is that they have adopted a long-term policy goal that no-one should be killed or seriously injured in a crash on their roads, "Vision Zero" as it is commonly called in Sweden, one of the pioneers (see Box 1). This forms the fundamental motivation behind the "Safe System" approach to delivering integrated road safety policies that bring road safety performance closer to that ultimate aspiration of innocuity.

Box 1. **Swedish approach to Safe System**

The 1980s in Sweden was a period when traffic safety stagnated. In the early 1990s the situation called for action and in the mid-1990s a new team in the Swedish Road Administration developed an innovative traffic safety policy that came to be called Vision Zero. The team was led by Claes Tingvall, the Traffic Safety Director at the time.

The Vision Zero policy was accepted at an early stage by the Swedish Transport Minister Ines Uusmann. Under her portfolio Vision Zero was taken to the Swedish Parliament in October 1997, and adopted as the new Swedish traffic safety strategy. The Parliament decided that the long term target should zero deaths or severely injured people in the road traffic system. The Parliament further decided that the design and function of the road transport system should be adapted to the demands coming from the Vision Zero strategy.

Vision Zero contained a shift in focus in many traffic safety areas. One important and significant shift was in the responsibility balance between the road users and system designers. System designers were defined as the bodies in society that design, operate and use the road transport system. Vision Zero is stating that the system should be adapted to the failing human. This constitutes a relatively dramatic shift from the common approach that road users should take the burden of a non-error tolerant road traffic system, summarised as:

- The designers of the system are always ultimately responsible for the design, operation and use of the road transport system and thereby responsible for the level of safety within the entire system.
- Road-users are responsible for following the rules for the safe use of the road transport system set by the system designers.
- If road-users fail to obey these rules due to lack of knowledge, acceptance or ability, or if injuries still occur, the system designers are required to take necessary further steps to counteract people being killed or seriously injured.

Vision Zero further focuses the road traffic safety challenge to the most severe and impairing injuries and fatalities. This was a change in the Swedish Road Administration that up until Vision Zero mainly used crashes as the key target. Shifting focus and targets from crashes to the most severe cases changed which solutions were prioritised. The safety core of Vision Zero is very much to design for the failing – non-perfect – human, i.e. recognising humans making misjudgements errors and mistakes.

The inception of the Safe System approach has parallels to the discussion around road safety as it relates to automated driving. A major component of early road safety policies in high-income countries in the 1950s

and 1960s was the assumption that the primary goal of road safety policy was to correct human errors in road crashes. This is somewhat mirrored in discussions around the safety-improvement potential of automated driving which putatively can eliminate most human errors or misjudgements in the driving task (discussed further).

The historic focus on driver error failed to acknowledge that inherent risks in road infrastructure and system design also were a significant contributor to crashes. Approaches to road safety built on that premise had limited outcomes and failed to significantly reduce the number of serious road injuries and deaths.

By the end of the 1970s, high-income countries started to implement some other elements: speed limits, compulsory seat belts and helmets, new infrastructure design, and expansion of the motorway network (the safest category of roads). This resulted in initial large reductions in road deaths but was then followed by a slowing in the rate of improvement and then later a levelling-off. Based on "blame the victim" attitudes where considerable focus and attention was given to improving the behaviour of the human being, road safety policies lacked the holistic approach needed to achieve further significant injury reduction.

In contrast, a Safe System promotes a wide combination of interventions, including stronger enforcement, safer road and roadside design and improved vehicle technologies, as well as better post-crash response. A Safe System does not view road deaths and injuries as the inevitable price to pay for a highly-motorised society. By seeing any road death as an unacceptable system failure, it counters the risk that transport planners may adopt measures of transport efficiency that tolerate fatalities that are affordably preventable.

Principles of the Safe System approach

In road safety analysis and crash studies, two approaches are possible. The traditional approach takes a backward-looking perspective. Standard crash causation analysis strives to understand all the factors involved in a crash in order to suggest ways how such a crash could have been prevented. This forensic approach has its limits for driver-involved crashes but one of its merits is that it builds a deep body of knowledge and experience regarding specific crash causation and contributory factors. This experience is lacking at present regarding automated driving and it isn't clear if the rate with which it can be collected can keep pace with technological developments – especially as many of these will no longer be linked to hardware but increasingly will involve rapidly changing code and algorithms embedded in software.

Alternatively, a forward-looking view considers what crashes might potentially happen in the future and identifies all possible ways for such crashes to be prevented. This proactive approach is the basis of the Safe System or Vision Zero approach for road safety (Table 1).

Table 1: **Comparing the traditional road safety approach and a Safe System**

	Traditional road safety policy	Safe System
What is the problem?	Try to prevent all crashes.	Prevent crashes from resulting in fatal and serious casualties.
What is the appropriate goal?	Reduce the number of fatalities and serious injuries.	Zero fatalities and serious injuries.
What are the major planning approaches?	Reactive to incidents. Incremental approach to reduce the problem.	Proactively target and treat risk. Systematic approach to build a safe road system.
What causes the problem?	Non-compliant road users or usage.	People make mistakes and people are physically fragile/vulnerable in crashes. Varying quality and design of infrastructure and operating speeds provides inconsistent guidance to users about what is safe use behaviour.
Who is ultimately responsible?	Individual road users.	Shared responsibility by individuals with system designers

How does the system work?	Is composed of isolated Interventions.	Different elements of a Safe System combine to produce a summary effect greater than the sum of the individual treatments- so that if one part of the system fails others parts provide protection.

Source: Inspired from New Zealand Transport Agency and VicRoads.

The following four principles underpin the Safe System approach:

- People make *mistakes* that can lead to road crashes.
- The human body has a *limited physical ability* to tolerate crash forces before harm occurs.
- A *shared responsibility* exists amongst those who design, build, manage and use roads and vehicles and provide post-crash care to prevent crashes resulting in serious injury or death.
- All parts of the system must be *strengthened* to multiply their effects; and if one part fails, road users are still protected.

Human error

Much of the focus on the potential safety benefits of automated driving have been centred on the elimination of human error in the driving task. That is because humans make mistakes in judgement, may drive impaired or distracted, may simply not be adequately aware of the driving environment or may not react quickly enough to rapid or unexpected changes. The Safe System approach inherently recognises capabilities and limitations of humans when designing and operating road transport systems.

Errors arising from interaction with the traffic and road environment can be limited by understanding these interactions and designing the road transport system from these interactions, in order to guide the road user to behave in a way that is as safe as possible. Yet, as human error cannot be fully eradicated there is a need, at the same time, to mitigate the consequences of mistakes. In simple terms, this basic principle of a Safe System starts with the insight that human error should no longer be seen as the primary cause of crashes. Instead, road crashes are seen as a consequence of latent failures created by decisions and actions within the broader organisational, social or political system which establishes the context in which road users act.

Limited physical crash tolerance

The human body has a limited physical ability to safely absorb the kinetic energy a crash exerts. If fault-free traffic performance cannot be ensured (e.g. there is no certainty that a crash can be avoided), the Safe System is designed so that speed differentials between potential crash opponents are reduced to levels that result in no harmful release of energy, by changing the material properties of vehicular or other surfaces that may enter into contact with each other so that they absorb crash-related kinetic energy to safe levels or by ensuring that crashes are materially impossible via robust separation techniques. Reduced differential operating speeds not only limits the harmful release of kinetic energy but it also limits the risk of errors and slow reaction times (ITF, 2018).

Contrary to this approach, today's road transport system has largely not been designed with the principle of mitigating common human error or compensating for it. For example, increased numbers of vehicles and higher travel speeds (often accompanied with smoother road surfacing) have negative consequences for road safety that have often overwhelmed efforts to improve the safety of road infrastructure, thus producing, on balance, reduced levels of safety.

Shared responsibility for road safety

There is a shared responsibility amongst those who design, build, manage and use roads and vehicles and provide post-crash care to prevent crashes resulting in serious injury or death. While it is the individual

responsibility of every road user to abide by safety-related laws and regulations (with education and enforcement being important factors to induce such behaviour), it remains a fact that human beings are not infallible and will always make mistakes, no matter how educated or law-abiding they may be.

In a Safe System, safe human behaviour in the first instance is informed and guided by the design, layout and operation of the road network, in addition to traditional education and enforcement actions for safe behaviour. Road designs and operations that provide feedback to users or are "self-explaining" can help create an environment that prompts safe road use.

In a system where user and usage mistakes are compensated for in a way that ensures no serious or fatal injuries, a large share of the responsibility for safety automatically shifts from the road users themselves to those who design the road transport system. These include road managers, the automotive industry, the police, transport operators, health services, the judicial system, schools, road safety organisations and, not least, politicians and legislative bodies. All of these bear joint responsibility for providing a road environment that increasingly anticipates potential mistakes and deals with them in a way that avoids serious harm.

Strengthen all parts of the system

The fourth principle underlying a Safe System addresses the potential outcome that if one element fails, serious injury may occur. In isolation, latent errors may not result in dramatic consequences. Indeed, these errors or design flaws may never manifest themselves. In those circumstances where they are triggered or activated, they may contribute to a chain of events that leads to a crash (Wegman and Aarts, 2006). To counter this, a Safe System strengthens all dimensions of road safety so that potential failures in one area are compensated for by other elements of the overall system design. The dynamic interaction effects amongst Safe System elements results in overall safety levels than are greater than the sum of each element's contribution to improved safety.

To provide a greater overall effect, the layers that together build a Safe System – the design and operation of road infrastructure, operating speeds, vehicles, human behaviour – are managed holistically rather than in separate silos. In those countries at the forefront of Safe System thinking, the four guiding principles are translated into concrete design principles for safety across the system safety, rather than for each component individually. This is a main difference with the traditional approach in which responses are often managed and implemented by different agencies.

The Safe System and automated driving

At its core, the Safe System addresses human fallibility and seeks to accommodate this within the traffic system so that mistakes don't lead to deaths or serious injuries. Automated driving promises a future where humans are (largely or completely) out of the driving loop and as such may seem to suggest that the foundational premise of the Safe System – human fallibility – is no longer an issue in achieving safe traffic performance. Traffic automatically becomes safer as humans are taken out of the driving seat. This assumption has largely driven the discourse around the push for increased vehicle automation on both the industry and policy side. It is not clear, however, that human-free driving will be safe. Nor is it clear that automated driving will be safer than conventional driving in every context though there are strong reasons to believe that it will deliver better safety outcomes in some cases. Part of this uncertainty stems from the original premise – that human errors are linked to over 90% of all fatal crashes.

Does eliminating driver error lead to safer outcomes?

Foremost among the predicted benefits of increased and complete vehicle automation is the promise of improved safety. With few exceptions (Shoettle and Sivak, 2015; Noy, et al., 2018) this predicted benefit is accepted uncritically on the basis of the observed rate of human error-involved crashes. Numerous studies

undertaking post-crash analysis of contributory factors have indicated that the vast majority of fatal vehicular crashes are linked to driver error – upwards of 90% of all fatal crashes (Sabey and Staughton, 1975; Hendricks, et al., 2001; Otte et al., 2009; Singh, 2015). The United States National Highway Transportation Safety Administration (NHTSA) indicates that 94% of fatal crashes in 2015 were associated with one or several human errors (NHTSA, 2016). The recent SHRP2 comprehensive naturalistic driving study in the United States looking at machine-logged vehicle activity and over 1000 crashes similarly found that human error contributed to nearly 90% of all recorded crashes (Dingus et al., 2016).

Faced with the consistency and the scale of these findings, many have argued that removing humans from the driving seat would lead to a 90% drop in crashes and significantly improve safety outcomes. This is both an untested and uncertain hypothesis - and certainly one that merits closer attention in light of the Safe System approach.

Clearly, the potential for automated vehicles to remove common and pernicious human errors and misjudgements from the driving task is significant (Anderson et al., 2016; Fagnant and Kockelman, 2015). Automated vehicles are not subject to being driven-impaired, being driven while texting or subject to other forms of human distraction or being driven while fatigued (respectively factors in 41%, 10% and 2.5% of fatal crashes in the United States (NHTSA, 2011; USBTS, 2014; USDOT, 2015)). However, it is reductionist to believe that human error has been properly identified as a contributory factor by those responsible for post-crash forensic investigation or that all crashes involving human error could have been otherwise avoided by addressing that error.

Human error does not imply driver responsibility

"Human error" is a specific concept used in forensic crash analysis to help classify and attribute operator-induced causes. It is based on the judgement of post-crash investigators, who exhibit varying levels of skill and experience, at the scene of the crash or on the basis of post-crash reports – themselves written by investigators with variable skills and experience. With the exception of impaired driving and time-stamped interactions with potentially distracting devices, such as phones, human error is the result of a deduction based on the absence of mechanical failure or infrastructure defects (Noy et al., 2018).

Despite new sources of incident-relevant data, differences in methodological approaches and other factors contribute to wide divergences in findings of human error contribution to crashes. One example is the case of fatigue as a contributory factor in crashes with involvement rates ranging from less than 5% to nearly 40% (Noy et al., 2018; Shinar, 2017). Thus a first element to consider when looking at the potential safety benefits of automation is that the reporting of "human error" involvement in fatal crashes may be overstated.

A second aspect to consider when assessing the scope for automation to improve safety outcomes by removing human errors in crash causation is that it does not follow that all crashes attributed to human error could have been reasonably avoided by drivers (Noy et al., 2018). Many crashes that involve human error also involve other factors that may have still led to a crash even if the human had not committed an error in judgement or misperception. Errors linked to poor roadway design (e.g. roads designed for lower speeds than legally allowed, confusing junction design, etc.) or faulty vehicle and interface design (confusing display or interfaces or visual obstruction) are often attributed to human causes when they are, in fact, design-induced errors (Noy et al., 2018). Human error can also be non-driver-related errors, by pedestrians, cyclists and motorcyclists. Since they won't be automated, their errors will probably not be eliminated by automation.

These considerations do not likely impact the general finding that automation may contribute to significantly better safety outcomes, but it may temper the assessment of automation benefits versus disbenefits. More generally, they indicate that the starting point for the discussion around the safety benefits of automation may not lie where many believe it to – namely that automation will improve road

safety outcomes by (more than) 90%. How *much* automation will improve road safety ultimately depends on how safely automated driving systems can carry out the parts of the driving task they are assigned. The technical skill with which these systems are able to handle the driver task without errors, glitches or unintended outcomes will matter here.

From the Safe System perspective, can we expect that removing human fallibility from the traffic equation will *de facto* lead to safer outcomes? Or will a new class of machine "fallibility" simply come into play. All that can be said with a great deal of certainty at this stage is that deploying automated driving systems that are designed such that they do not replicate human errors or risky behaviour is that the result will be an elimination of those human errors and risky behaviours in the driving task. This is not the same thing as saying that automated driving will be safe – or safe enough – though it is plausible that it will be safer than human driving in many instances. Machines will crash and thus the central tenets of the Safe System should still hold in the design of the road traffic system and environment. In the case of automated driving, the object of the approach switches from human mistakes to programming errors or unwanted outcomes from automated driving systems.

Our collective understanding of these matters is still at an early stage since experience with these technologies is relatively low and the technological systems are themselves evolving at a very rapid pace. A full application of Safe System thinking must consider building the transport system in a way to accommodate failing automated vehicles. The design parameters related to speed, design etc. will be similar to the ones for failing humans. Aiming for perfection in automated driving systems is important, but a Safe System should remain the fall-back solution. A system built to safely absorb human error is also tolerant to machine errors. In the next section, we address the challenge of adequately assessing the safety performance of automated driving and explore where the Safe System approach may provide a helpful framework for addressing uncertainty in this domain.

Assessing the safety benefits and disbenefits of automated driving

Automated driving systems that either assist or replace humans in the driving task are rapidly being designed, tested and, in many cases, deployed in pilots around the world. Understanding the safety implications and performance of these systems is complicated by the lack of historic data and the diversity of technologies and design missions. It is further complicated by the lack of a common framework for assessing safety and by rapid changes in the object of assessment – a combined hardware and software system where upgrades in code can dramatically change the performance characteristics of the system. Finally, it is linked to the skill with which the combination of hardware (sensors, processors and actuators) and software mimics or improves on the driving performance of humans. All of these have implications for the relevance of the Safe System approach to automated driving.

At the outset is perhaps a fundamental question that is rarely fully articulated – why automate in the first place? (Noy et al., 2018) According to (Wickens et al., 2003) there are a range of potential motivations for seeking to automate human functions including:

- When it is *dangerous* for humans to carry out a task (e.g. driving impaired).
- When it is *impossible* for humans to carry out a task (e.g. accurate night-time sensing or sensing beyond line of sight).
- When carrying out a task is *difficult* for humans (e.g. rapidly reacting to sudden obstacles).
- Just *for the sake* of automation (e.g. as a way of creating a market even if no safety benefit is conferred to users).

The current development of automated driving systems is motivated by all of these in different measure.

Perhaps a more safety-relevant way of addressing the issue of what motivates automation centres are the questions: What tasks can automated systems do better than humans (from a safety perspective)? And, what tasks can humans perform better than automated systems? These are central questions that have accompanied many past technology developments and was articulated by (Fitts, 1951) in a framework that has remained surprisingly relevant today even in the context of vehicle automation and technologies that were barely imaginable at the time (Cummings, 2014; de Winter and Dodou, 2014).

Table 2 shows the updated "Fitts List" which addresses ideal function allocation between hardware/software systems like automated vehicles and humans (Schoettle, 2017). One of the key premises of Fitts still holds today in light of driving functionality – tasks that humans are better at than automated driving systems should be performed by humans and tasks that automated driving systems are better at performing than humans should be performed by automated driving systems. Where there is conditionality – e.g. better driving performance for either humans or automated driving systems is linked to a specific set of conditions or contexts - the Safe System approach implies that the resulting ambiguity does not lead to crashes, loss of life or serious injuries. This may entail upstream *system* versus *vehicle* design that seeks to eliminate these risks.

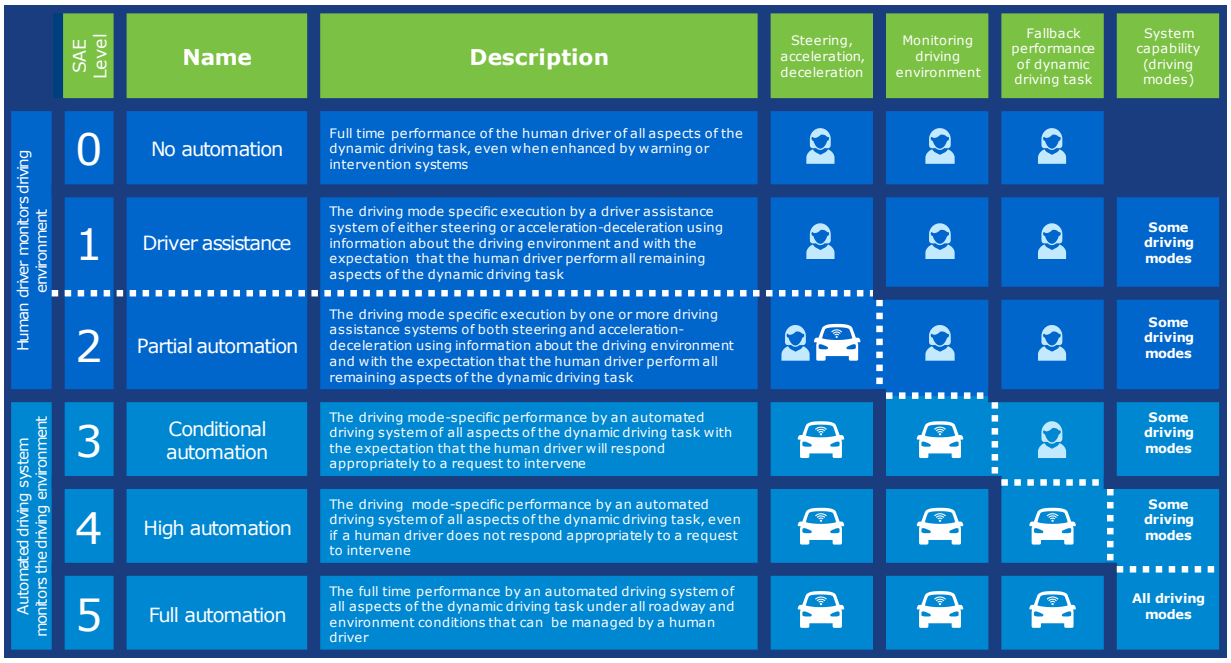
There are three broad potential crash causation scenarios for automated vehicles. The first encompasses all those situations where faulty human-machine interactions contribute to a crash. These interactions are only possible in systems that expressly allow this interaction – which is itself a hotly debated design issue. The second is when sensors do not detect a critical part of the vehicle environment or incorrectly identify it. Sensor capabilities have increased tremendously in recent years but, alone or in combination, they still have weaknesses that compromise safe system performance in certain contexts. A final category of automated vehicle crashes is those where the automated driving system encounters a situation unforeseen and unanticipated by its code and algorithms (or by its artificial intelligence engines). In these instances, the resulting actions – including the decision to not act – can lead to crashes. We discuss these in the sections that follow.

Table 2: **Summary of Fitts List of strengths and weaknesses across various aspects of function allocation between humans and hardware/software systems**

Aspect	Human	Hardware/Software system
Speed	Relatively slow.	Fast.
Power output	Relatively weak, variable control.	High power, smooth and accurate control.
Consistency	Variable, fatigue plays a role, especially for highly repetitive and routine tasks.	Highly consistent and repeatable, especially for tasks requiring constant vigilance.
Information processing	Generally single channel.	Multichannel, simultaneous operations.
Memory	Best for recalling/understanding principles and strategies, with flexibility and creativity when needed, high long-term memory capacity.	Best for precise, formal information recall, and for information requiring restricted access, high short-term memory capacity, ability to erase information after use.
Reasoning	Inductive and handles ambiguity well, relatively easy to teach, slow but accurate results, with good error correction ability.	Deductive and does not handle ambiguity well, potentially difficult or slow to program, fast and accurate results, with poor error correction ability.
Sensing	Large, dynamic ranges for each sense, multifunction, able to apply judgement, especially to complex or ambiguous patterns.	Superior at measuring or quantifying signals, poor pattern recognition (especially for complex and/or ambiguous patterns), able to detect stimuli beyond human sensing abilities (e.g., infrared).
Perception	Better at handling high variability or alternative interpretations,3 vulnerable to effects of signal noise or clutter.	Worse at handling high variability or alternative interpretations,3 also vulnerable to effects of signal noise or clutter.

Source: (Schoettle, 2017) adapted from (Cummings, 2014; de Winter and Dodou, 2014)

Figure 1. **The five levels of automated driving**



Source: Based on SAE Society of Automotive Engineers

With or without the driver?

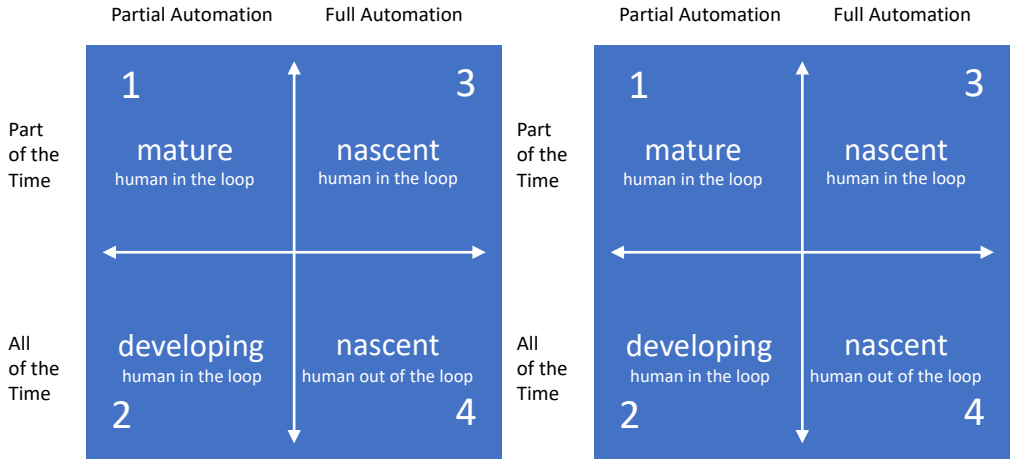
When assessing the safety implications of automated driving, it is important to bear in mind that not all automated systems share, or are targeting, the same performance. The capabilities of these systems vary from simply assisting drivers in certain contexts and in limited capacities to fully replacing the human driver in specific contexts. The former form the basis of a number of technologies already included in commercially available cars and trucks (e.g. lane assist, self-parking functions, limited autopilot function) whereas the latter only comprise vehicles that are being tested in various trials, many of them in the United States given the relatively permissive testing environment offered there by several states.

Figure 1 sets out the 5 levels of automated vehicle performance as defined by the International Society of Automotive Engineers (SAE, 2015). It categorises vehicle automation functions according to the level and scope of driving task responsibilities allocated to the human driver versus the automated system – or to both in some instances. A core issue with regards to safety is how well the handover from automated systems to human drivers occurs at SAE levels 2 and 3 when the system cannot interpret its environment satisfactorily. Machine-to-human handovers also are triggered at SAE level 4 when the driving context changes to one that is beyond the capabilities of the automated system (e.g. when passing from a simplified driving environment on a grade-separated roadway to a complex street environment).

The SAE levels help develop a complete taxonomy of automated driving system capabilities and functional boundaries but, fundamentally, only two dimensions matter (Noy et al., 2018; ITF, 2015) (Figure 2):

- Does the automated system seek to assist or replace the driver - is automation partial or complete?
- Does the automated system operate part of the time in some contexts or everywhere at all times?

Figure 2. **Two-dimensional categorisation of automated driving systems**



Source: Adapted from (Noy, et al.)

The ability for advanced driver assistance systems to offer partial automation part of the time is already quite advanced at present (e.g. quadrant 1: lane-keeping, automatic speed control) and the capability for systems to provide partial automation in all contexts is developing (e.g. quadrant 2: lane-keeping assist, autonomous cruise control and lane-change capability, etc.). The technologies that allow full automation part of the time and in certain contexts is nascent but developing (e.g. quadrant 3: motorway autopilot or traffic jam assist). A significant gap exists, however, between those systems and ones that could completely replace human drivers is all contexts and at all times. In quadrants 1 through 3 just as in SAE levels 1-3, humans retain a significant role as driver, back-up driver and/or supervisor of the automated system. This hybridisation of roles poses some inherent safety challenges.

Much has been said about the difficulty in ensuring safe driving performance in the presence of mixed driving task allocation – especially at SAE levels 2 and 3 and in quadrants 1 through 3 in figure 2. (Le Lann, 2017; Noy et al., 2018; Ni and Leung, 2016; Reschka, 2016). This difficulty stems from keeping humans "in the loop" and is related to a number of "ironies of automation" as noted in (Noy et al., 2018). These are termed "ironies" because far from alleviating the driving task, the hybrid allocation of driving to both automated systems and humans may in fact lead to more complex decision-making environments and increased risks of unintended and safety-reducing consequences.

At their core is the finding that as automation increases (up to the point where humans are completely removed from the driving task) the more drivers are challenged under the most critical circumstances. The principal "ironies" described by (Noy et al., 2018) (Bainbridge 1983) are:

- **Task allocation**: Poorly adapted task allocation occurs when easy tasks are allocated to automated driving systems (tasks which average human drivers generally handle well) leaving only the most complicated and cognitively difficult tasks to the human driver. This increases the potential for errors or unsafe outcomes – automation should target tasks that are impossible for humans to do well under certain circumstances.
- **De-skilling** (or automation addiction – the term used by the US Federal Aviation Administration (FAA) to describe pilots whose flying skills diminish over time in the presence of automated flying systems): Lack of practice or imperfect situational awareness leads to reduced skill and delays by humans in carrying out driving functions when they are required to do so by disengagement of the automated driving system. A secondary consideration is that the impacts of de-skilling are largely unknown on a heterogeneous driving population with varying capabilities and levels of training.

This a particularly important factor to consider under higher levels of automation where drivers spend less time in effective control of the vehicle.

- **Cognition:** Lack of cognitive engagement in the driving task leads to lower levels of situational awareness and longer reaction times when the automated driving function disengages. Simply supervising the operation of the automated driving function does not confer sufficient cognitive engagement to prevent loss of vigilance. A secondary factor is that non-engaged drivers easily become bored and may engage in other distracting activities that also limit the speed and effectiveness of system handovers.
- **Control:** Issues of cognition relate to the *understanding* of the driving context at the time of a handover from the automated driving system, but driving is also a *learned skill* that must be practised to be perfected. Less time spent driving and less recall of the physical "feel" of vehicle response characteristics can lead to unsafe driving response in the form of poorly modulated steering, acceleration or deceleration.

These weaknesses confound efforts to ensure the safe operation of those automated driving systems that are designed to share driving tasks between humans and drivers. Design workarounds can include warning systems, targeted training and situationally-adaptive handover protocols (e.g. a system sensing driver fatigue will trigger earlier handoffs than for drivers it senses are alert). The extent to which these workarounds and system design changes adequately address the challenges of poorly adapted task allocation, de-skilling, cognition and control issues in handover protocols is unclear. This is partially due to the lack of statistically valid data on the efficacy of these measures (discussed further) but it may be that these are intractable vulnerabilities. If that is the case – or even if it is the case that we simply do not know if these issues can be adequately compensated-for in the design of automated driving system – the Safe System approach suggests two alternative outcomes:

The first is that these challenging machine-to-human handovers should be avoided entirely. Some have suggested that SAE level 3 automated driving is inherently unsafe and perhaps level 4 as well. Several players in the automated driving field share this vision and are designing systems that completely jump over intermediate levels of automation thus avoiding the potential for crashes that result as human operators take over from automated systems.

The second outcome would be that, accounting for the inherent risk of machine-to-human handovers, these should only occur in such a way that faulty handovers do not result in death or serious injury. This may suggest that these handovers should only take place at speeds where crashes will not lead to serious fatalities or deaths or in simplified traffic contexts where predictability of the handover is high and the potential for a faulty handover is extremely low. In the former case – slow speed operation – humans may drive as safely as automated systems thus partially negating the attraction of those use cases.

Generally, however, hybrid, "human in the loop" automation scenarios seem to challenge some of the core principals of the Safe System approach. This is a serious consideration for authorities that otherwise are seeking to deliver on Vision Zero and design safe traffic systems.

Box 2. **Principles for integrating Automated Vehicles into Boston's Vision Zero approach**

Cities can and should use their Vision Zero programs and associated funding to embark on some or all of the following, as Boston has begun to do:

- Employ pilot projects for fully autonomous vehicle technologies at gradually larger geographic scales, and measure the before and after safety and network efficiency outcomes. Restrict or ban human-controlled driving from certain districts as necessary to control the environment further and illustrate potential benefits of a completely driverless city. Urban islands, pedestrian-dominated districts, and campuses (academic, corporate, or otherwise) are logical places to start.
- Apply camera and sensor technology to automate enforcement, collect collision and "near miss" information, and perform before and after studies of any pilot projects of AV technology. If legislative approval is required to do so, discerningly tie the automated data collection program in with pressing and much needed Vision Zero safety goals.
- Use data to make the case for continued AV rollout, in the context of Vision Zero. In the same way that a street redesign project can be shown to improve safety outcomes, do the same detailed before and after measurement for early-stage autonomous vehicle pilot projects. Tell the story to clearly illustrate any safety benefits, and demonstrate how AVs can help reach Vision Zero.
- Reference Vision Zero consistently on all AV policy development and strategic decisions. This will aid in public understanding of an abstract issue, in gaining political capital to test things out, and potentially in the ability to capture federal or other grant funding for AV projects.
- Adopt a short- and long-term AV strategy as part of an updated or new Vision Zero Action Plan. Each city is unique and will face specific implementation challenges on AVs that require a proactive approach to grow local AV market share. Include specific goals, strategies, and benchmarks that lay out rationale, policy levers, and measurement tools that will be applied to execute an AV strategy in the context of Vision Zero.

Source: City of Boston (www.boston.gov/departments/new-urban-mechanics/autonomous-vehicles-bostons-approach)

Comparative skill in sensing the road environment: Human drivers and highly automated driving systems

At the core of all levels of automation is the ability for automated vehicles to perceive ("sense") their environment, process this information to determine what is relevant to safely, carry out the allocated driving task, decide on a course of action, successfully carry out this action by triggering actuating systems (e.g. steering, braking, signalling, etc.) and then assess the result of the action carried out (Parasuraman et al., 2000). Each of these steps mimics how humans drive but with different capabilities and speeds. Until recently, the skill with which the combined sensing, processing, deciding, acting and assessment cycle could be carried out by automated systems has been below what humans could achieve both in scope and latency.

This is no longer the case, at least not in many contexts. Technology is now reaching the point where the fusion of different sensors, processing systems and actuators can replicate and in some cases improve on human driving performance. This convergence is at the heart of the incipient revolution promised by highly- and fully-automated driving. But the convergence is not complete and there remain key areas where automated driving systems still lag behind the capabilities of the average human driver. Further, beyond correct perception and decision-making functions, the issue of validating these systems remains challenging as discussed further (Stolte et al., 2016).

Table 3: **Assessment of various driving-related sensing systems**

Eyes (Human drivers)	Colour, stereo vision with depth perception. Large dynamic range. Wide field of view, moveable both horizontally and vertically. Range: No specific distance limit; realistic daytime limit of at least 1000 metres and realistic night-time limit of 75 metres under low-beam headlamp illumination.
Radar (Automated vehicles)	Accurate distance assessment. Relatively long range. Robust in most weather conditions. Immune to effects of illumination or darkness. Fixed aim and field of view (deploying multiple radar sensors can compensate for this). Field of view (horizontal): ~ 15° (long range) to ~ 90° (short range). Range: ~ 250 m. Resolution: ~ 0.5° to ~ 5°.
Lidar (Automated vehicles)	Accurate distance and size information. Able to discern high level of detail (shape, size, etc.), especially for nearby objects and lane markings. Useful for both object detection and roadway mapping. Immune to effects of illumination or darkness. Fixed aim and field of view, but able to employ multiple lidar sensors as needed (although some lidar systems are capable of 360° within a single piece of equipment). Field of view (horizontal): 360° (maximum). Range: ~ 200 m. Resolution: ~ 0.1°.
Camera systems (Automated vehicles)	Colour vision possible (important for sign and traffic signal recognition). Stereo vision when using a stereo, 3D, or time-of-flight (TOF) camera system. Fixed aim and field of view, but able to employ multiple cameras as needed. Field of view (horizontal): ~ 45° to ~ 90°. Range: No specific distance limit (mainly limited by an object's contrast, projected size on the camera sensor, and camera focal length), but realistic operating ranges of ~ 150 m for monocular systems and ~ 100 m (or less) for stereo systems are reasonable approximations. Resolution: Large differences across different camera types and applications.
Dedicated short-range communications – DSRC (Connected vehicles)	Applicable to vehicles operating at any automation level. No line-of-sight requirement (omnidirectional antenna). Robust in weather conditions. Able to both receive and send detailed information. Range: Long range (~ 500 m) that can be effectively extended by communicating with transportation infrastructure in addition to other vehicles; however, the signal strength of transmissions decreases based on the inverse-square law (i.e., signal strength is inversely proportional to the square of the distance from the transmitter). Can communicate future actions or planned manoeuvres (especially for AVs) to other traffic, alleviating need for other traffic to sense and/or predict what the connected vehicle will do . Can communicate information about recently encountered roadway conditions, traffic conditions, etc. to other roadway users. Able to communicate with other road users or transportation modes within the interconnected DSRC system (e.g., pedestrians, trains, etc.).

Source: (Shoettle, 2017)

Automated driving systems mobilise a number of sensing devices to develop a sufficiently clear picture of the driving environment on which to make correct and safe decisions. These sensing capabilities include different forms of embarked technologies and can be complimented by other inputs that help the automated driving system assess its environment (e.g. GNSS location data, high definition mapping, information potentially communicated from other vehicles or from infrastructure). As vehicles move up to higher levels of automation on the SAE scale, the relative importance of these sensing inputs increases as the potential for human drivers to make corrective action (using their own sensing and cognitive capabilities) diminishes and, ultimately, disappears.

Table 3 describes and assesses the principal performance aspects and relative advantages of various sensor technologies in comparison to human eyesight. Each of these human or machine sensing platforms must be able to, singly or in combination with other sensors, adequately replace or improve on human vision and perception if they are to improve safety outcomes. Their capacity to do so, however, will be challenged in a number of situations (Shoettle, 2017):

- Extreme weather or other degraded environmental conditions such as heavy rain, snow or fog). These phenomena reduce maximum range and signal quality for human vision, optical sensors (cameras, lidar) and DSRC – although to a lesser extent.
- Excessive dirt or physical obstructions such as snow or ice on the sensor surface or the vehicle surface. These phenomena reduce maximum range and signal quality for human vision, cameras, lidar and radar.
- Darkness, low illumination or glare. These phenomena reduce maximum range and signal quality for human vision and cameras.
- Large physical obstructions (buildings, terrain, heavy vegetation, etc.) interfere with line of sight for human vision and all basic AV sensors (cameras, radar, lidar); some obstructions can also reduce the maximum signal range for DSRC.
- Dense traffic: Interferes with or reduces line of sight for human vision and all basic AV sensors (cameras, radar, lidar); can also interfere with effective DSRC transmission caused by excessive volumes of signals/messages. However, human drivers do have some limited ability to see through the windows of adjacent vehicles.

Table 4: **Assessment of sensor performance across driving tasks**

Performance aspect	Human	Automated Vehicle			Connected vehicle	Connected, automated vehicle
	Eyes	Radar	Lidar	Camera	DSRC	Radar, Lidar, Camera and DSCRC
Object detection	Good	Good	Good	Fair	n/a	Good
Object classification	Good	Poor	Fair	Good	n/a	Good
Distance estimation	Fair	Good	Good	Fair	Good	Good
Edge detection	Good	Poor	Good	Good	n/a	Good
Lane tracking	Good	Poor	Poor	Good	n/a	Good
Visibility range	Good	Good	Fair	Fair	Good	Good
Poor weather performance	Fair	Good	Fair	Poor	Good	Good
Dark or low illumination performance	Poor	Good	Good	Fair	n/a	Good
Ability to communicate with other traffic or infrastructure	Poor	n/a	n/a	n/a	Good	Good

Source: (Shoettle, 2017)

On the basis of a comprehensive review of human and machine sensing capabilities applied in the case of various driving and exemplar pre-crash scenarios, researchers at the University of Michigan found mixed results regarding the ability for hardware/software systems to replicate and improve on human sensing capabilities (Shoettle, 2017). Table 4 summarises their assessment of how different sensor strategies can handle individual driving tasks. They found that machine sensing is generally well-suited for performing driving tasks in terms of reaction times, consistency and multichannel information processing. Some sensing technologies allow supra-human sensing capacities (e.g. infrared sensing, accurate distance measurement, seeing through the dark or inclement meteorological conditions). At slow speeds and in degraded conditions, the performance of automated driving systems may exceed human performance exercised under ideal conditions.

It should be noted that the performance of DSRC-based sensing described by (Shoettle, 2017) in Table 4 is conditioned on the type of information that would be included in the messages transmitted. As currently envisaged, messages that include vehicle identity and location would allow DSRC performance for "object

detection", "object classification" and "lane tracking" to be as good or better than humans if messages are transmitted and received in time.

Humans still retain an advantage over single sensor-based automated systems when it comes to reasoning and anticipation, perception and sensing when driving. Overcoming this gap (and only in certain conditions) requires multi-sensor fusion on the part of the automated system. This strategy is commonly employed on various vehicle testbeds deployed in current trials. Even in the case of multiple sensor fusion, human capabilities still outperform that of automated systems in certain problematic and complex contexts. Some common traffic scenarios still confound automated driving system capabilities. These include straight crossing path and left turn across, opposite direction crashes which are consistently difficult for automated systems to assess and avoid (Shoettle, 2017; NHTSA, 2017)

The risk stemming from these and other dangerous scenarios can potentially be mitigated by augmenting embarked sensing capabilities with inputs from other vehicles and infrastructure (as in the case of information relayed by DSRC). The need to move from a "reactive" safety paradigm where vehicles rely solely on their embarked capabilities to a "proactive" safety framework where vehicles are embedded in a communicative network to deliver better safety outcomes is actively debated. Automated systems that can "see" what humans cannot (e.g. beyond line of sight) and relay this information to each other show promise for surpassing human driving capabilities. But the communicative car strategy is one that is not void of new risks and challenges.

The first of these regards the technical proficiency with which automated driving systems can and need to address other vehicles within and outside of their line of sight. There are a number of technical challenges that have historically stymied the effectiveness of this strategy. These include the propensity for dropped and failed messages as a result of both their volume and weaknesses inherent to the channels used to communicate these. In high traffic environments, safety critical vehicle positioning messages may arrive simultaneously leading one to be discarded or ignored (collision exclusion), they may be lost in transmission, they may contain errors or may be incompatible with other received messages due to latency in transmission and processing. Many of these issues can be addressed via message-handling protocols but they increase the computing cost of these systems and introduce delays that may impact accurate context awareness.

These limitations were partly a result of early vehicle-to-vehicle communications standards, e.g. IEEE 802.11p, being largely based on Wi-Fi standards. (Wu, et al. 2013) note that "Vehicular communication environments differ significantly from the sparse and low-velocity nomadic use cases of a typical Wi-Fi deployment. Thus, there are many challenges to adapt Wi-Fi technologies to support the unique requirements of vehicular communications such as achieving high and reliable performance in highly mobile, often densely populated, and frequently non-line-of-sight environments".

Significant progress has been made in addressing these shortcomings, most notably with the European Telecommunications Standards Institute ITS-G5 standard. This standard is based on IEEE 802.11p which has been adapted and optimised to handle the high velocity and dynamic automotive environment. Other standards are under development that leverage 4G and 5G cellular networks that seek to address some of the shortcomings of the ITS-G5 standard (HÄRRI et Berens 2017) (Eckhoff, Sofra et German 2013) but the ITS-G5 is already available for deployment in support of highly automated driving. Though these standards are being tested, some extensively, in live and virtual settings, there is still little experience regarding their performance in fully scaled up and complex traffic applications – especially as concerns safety-critical messaging.

More fundamentally, there is a real question as to the usefulness of full awareness of all potential vehicles and infrastructure elements within the extended sensing range of cooperative and communicative vehicle/infrastructure networks. In the connected vehicle ecosystem approach advocated by many

automotive and ITS stakeholders, information on vehicle positioning and condition is broadcast from all vehicles to all vehicles. Each driver (or vehicle in the case of high levels of automation) can use this to create a dynamic situational map that extends beyond line of sight and which serves as an input to anticipatory driving decisions. Much of that information is superfluous to the immediate and proximate driving task, however. Parsing this information down to just those objects (vehicles and infrastructure) that are relevant to safe driving performance and collision avoidance may prove a more computing and data efficient strategy. This parsing can take place within each automated driving system on the basis of the information it receives from all other vehicles. Thus vehicles "perceive all" but decide to act only on what is near or critical. This strategy is at the core of the connected automated vehicle paradigm (C-ITS, 2016, 2017).

Another approach involves the creation of ad-hoc "cohorts" - instant, temporary and self-adjusting proximate networks that only track those vehicles and objects that could potentially impede safe driving and ranks them on their relative potential for driving task disruption (either by their proximity of because of their lane position) (Le Lann, 2017). This approach avoids some of the issues relating to messaging failures inherited from Wi-Fi-based standards. It also reduces the collection and storage of sensitive data regarding all vehicles in sensor/network range – information regarding the identity of the vehicle is not necessary under this approach, only its relative position to the reference automated vehicle.

A second approach that holds promise for handling how sensor fusion can improve automated driving outcomes is the Responsibility-Sensitive Safety (RSS) concept developed by MobilEye/Intel (Shalev-Shwartz, 2017 - see Box 3). This approach proposes a formal method of establishing automated driving behaviour that mimics heuristics used by human drivers. The RSS model is built on a mathematical model that accounts for sensor perception and driving system reaction latencies in order to define two constantly updated vehicle states:

- The *safe state* where there is no risk that the automated vehicle will cause a crash even if nearby vehicles make unpredictable or unsafe manoeuvres.
- Default *emergency policies* which comprise the most aggressive action that the automated vehicle can make and still maintain or return to the safe state.

These two states form the boundaries of a constantly changing safe-operating bubble that is the primary controlling variable in the automated driving task. They represent a mathematical framework that shows potential for embodying Safe System principals.

One factor to consider is that the driving behaviour of systems operating under the RSS framework, or any automated driving system hard-coded to avoid unsafe situations or potential crash risks will drive differently than human drivers. A focus of current trials is how to incorporate more naturalistic driving styles adapted to local contexts in a way that does not compromise safety performance. There may be a hard border to how far this is possible. This raises the question of the potential consumer acceptance and commercial feasibility of truly safe automated driving systems. It seems likely that these two factors are linked to the original intent of the automated driving strategy. In the first instance, if automated driving seeks to fully replicate human driving behaviour and uses, passengers may be wary of systems that they perceive to drive differently than they themselves do. If, on the other hand, commercial deployment pathways for automated driving focus on different use cases than the "own car driving" scenario, expectations and acceptance of safe driving behaviour may be greater. Much as passengers may wish to be driven by "safe" taxi drivers, they may also wish to be driven by "safe" automated driving systems. This may be a factor to consider when looking at the potential for automated fleet-based shared mobility services such as those announced by a number of companies.

The second challenge to integrating "network" sensing to automated driving platforms is addressed further on and relates to the cyber-security risks associated with opening up critical driving sub-systems to external communication and control vectors.

Box 3. **Responsibility-Sensitive Safety concept**

Responsibility-Sensitive Safety (RSS) is a safety concept for automated driving developed by Mobileye which formalises the common sense of human judgement with regard to the notion of "who is responsible for causing an crash". RSS is interpretable, explainable, and incorporates a sense of "responsibility" into the actions of a automated driving system. The definition of RSS is agnostic to the manner in which it is implemented — which is a key feature to facilitate the goal of creating a convincing global safety model. RSS is motivated by the observation that agents play a non-symmetrical role in a crash where typically only one of the agents is responsible for the crash and therefore is to be blamed for it. The RSS model also includes a formal treatment of "cautious driving" under limited sensing conditions where not all agents are always visible. The goal is to guarantee that an agent acting under RSS will never *cause* a crash, rather than to guarantee that an agent will never be involved in an crash (which may be impossible).

RSS is not a formalism of blame according to the law but instead it is a formalism of the common sense of human judgement. For example, if some other car violated the law by entering an intersection while having the red light signal, while the robotic car had the green light, but had time to stop before crashing into the other car, then the common sense of human judgement is that the automated car should brake in order to avoid the crash. In this case, the RSS model indeed requires the automated driving system to brake in order not to cause a crash, and if it fails to do so, it shares responsibility for the crash.

Supporting RSS is a "semantic" language that consists of units, measurements, and action space, and specification as to how they are incorporated into Planning, Sensing and Actuation of the automated driving system. This language formalises how humans assess the driving environment. Humans do not base their decisions on geometric notions — they do not think in terms of "drive 13.7 meters at the current speed and then accelerate at a rate of 0.8 m/s2 ". Instead, they deploy heuristics of a semantic nature — "follow the car in front of you" or "overtake that car on your left". The language of human driving policy is about longitudinal and lateral goals rather than through geometric units of acceleration vectors. This semantic model is crucial on multiple fronts connected to the computational complexity of planning that do not scale up exponentially with time and number of agents, to the manner in which a function that maps the "sensing state" to an action.

Adapted from: Shalev-Shwartz, 2017

Coding for the unexpected

As described above, human perception and deductive reasoning capabilities remain unchallenged when considering the full range of contexts and situations that could potentially be encountered during the driving task. Planning and accounting for the unexpected is a human strength that is difficult to replicate in automated driving systems. Part of the difficulty rests in the fact that machine behaviour must be formally represented by code and embedded in specific algorithms. Coding for automated driving has historically been built on formal "if-then" decision heuristics. These mimic those used by humans in many cases, but not necessarily in complex or unfamiliar situations (Dreany et al., 2018).

One complication related to coding for uncertainty is that there are no simple or complex algorithms that cover the full range of pre-crash and crash situations. This increases the number of algorithms necessary to address a wide range of potential safety-critical situations. However, as the volume of mission-critical code grows, so does the potential for errors and unexpected interaction effects. Code embedded in automotive systems are prone to errors (20-50 errors for every 1000 lines of code, 15% of which are missed by industry standard quality assessment techniques (Reader, 2018 – Fast Company; Lonsdale Systems, 2016)) and this raises the potential for errors that contribute to unsafe outcomes.

Unlike industry-standard hardware safety devices (seat belts, standard steering and braking controls, protective airbags, etc.) there are no standard safety algorithms that have been adopted or mandated. Indeed, algorithmic skill in managing the driving task is very much a point of competition amongst different

automated driving systems. If this should be the case, going forward is an open question and one that raises issues relating to intellectual property rights and automated driving system differentiation. However, if one set of algorithms leads to demonstrably safer outcomes than others, public authorities will have to assess how to ensure that these are embedded in all licensed automated driving systems - as to not do so would break with the Safe System approach and lead to poorer safety outcomes.

Simple and binary "if-then" decision trees also display path dependency which may lead to unexpected safety-reducing outcomes. A decision that may correctly result from an original "if-then" proposition may no longer be the correct decision outcome once the full chain of events and decisions are considered. This may result in a situation where "correct" actions as seen from the perspective of code lead to unwanted and unforeseen outcomes.

One way to move beyond these limitations in code is to deploy artificial intelligence (AI) frameworks that are capable of "self-learning" behaviour. These approaches move beyond "if-then" heuristics and mimic human decision-making in complex and unexpected situations by merging learning, memory and decision-making. How well they do so – especially in the face of unexpected situations – largely depends on the scale and scope of the learning they have undergone. Much as the quality of human decision-making builds with increased experience, so does the quality of AI with the amount and range of data they ingest. Nonetheless, AI systems can also be challenged when situations present themselves that are out of the range of what the AI is capable of handling (Le Lann, 2017).

Further, deep machine learning of the type employed by AI helps in augmenting on-board scene recognition capabilities. Better scene recognition putatively leads to better (safer) decisions on the part of the automated driving system. Yet, without inter-vehicle communications, these decisions are only used by the vehicle at hand and, eventually, other vehicles from the same manufacturer if these decisions are broadcast back to company servers. In the former case, AI only contributes to a particular vehicle's reactive safety capabilities and in the latter, AI-derived safety-relevant knowledge is only shared with some of the vehicles on the road.

Whether code is simple or complex, there will be situations that fall outside of the boundaries of what it can handle. In the context of the Safe System approach, this means that automated driving systems must be able to recover to a safe operational state (in this case, where the context reverts to one that the automated system code can handle). In contexts where complexity or uncertainty is likely to push code to its limits, the Safe System approach would dictate that vehicle operation should be proactively managed to parameters that meet imperatives for reduced potential kinetic energy from crashes or separation of potential crash opponents.

Evaluating crash experience of automated versus conventional vehicles

A final issue to address when discussing the applicability of the Safe System Approach to automated driving is the level with which the safety performance of automated driving systems can be reliably assessed. This is a crucial element to consider since public authorities have a mandate to license technologies and their uses on public roads in such a way that they do not lead to unsafe outcomes. This is problematic in the case of automated driving since experience with the safety performance of these systems is lacking, the object of safety regulation is rapidly and constantly evolving and there is no consensus on what level of safety should be guaranteed by automated driving systems.

Assessing and ensuring safety

Few statistically valid assessments of the safety performance of automated driving systems exist to-date (Kalra and Paddock, 2016; Noy et al., 2018). Part of the issue is the lack of appropriate metrics or robust benchmarks.

For example, reported crash incidents provide insight into those cases where an automated driving system was not able to avoid a collision with another traffic participant. This may seem as a valid metric to use in assessing the safety of automated driving. But nearly all automated driving systems currently being tested in high automation use cases (SAE level 4 and above) rely on back-up human drivers to recover control of the vehicle when the automated driving system reaches the limits of its performance capabilities. Automated driving system disengagements, and not just crashes, seem a more relevant metric to track safety performance since, absent the driver, the vehicle may have crashed. Further, if the ultimate goal is to track the skill with which automated driving systems can copy or improve on human driving, the appropriate comparison set includes the number of near-misses avoided by human drivers. Absent large scale naturalistic driving studies such as SHRP2, this metric is largely unavailable.

Some authorities require companies testing their vehicles to report on disengagements, but not all (e.g. California requires disengagement statistics reports whereas neighbouring Arizona does not). This potentially leads to a situation where competition for the least stringent testing environment emerges which may have an impact on the safety performance of vehicles tested on open roads in these jurisdictions.

A second factor to consider regarding on-board back-up drivers is that because of these disengagements, a significant share of driving by automation-capable vehicles in in fact carried out by the human driver. The number of kilometres driven in *automated mode* is more relevant than the number of kilometres driven overall by these vehicles.

Two recent studies have attempted to assess the comparative safety performance of automated driving on the basis of reported crashes per distance driven for automated versus conventionally driven cars – one by the University of Michigan Transportation Research Institute (UMTRI) (Shoettle and Sivak, 2015) and one by the Virginia Tech Transportation Research Institute (undertaken for Google) (Blanco et al., 2016). Both studies found that automated driving systems were characterised by much lower crash rates than conventionally driven cars irrespective of crash severity categories or reference conventional car populations. However, the very small number of automated driving system-involved crashes (none fatal, though this has now changed) make these results statistically insignificant. For this to change, automated driving miles travelled (and crash incidents reported) would have to scale up considerably.

The current approach largely employed in advanced automated driving system trials worldwide is to drive these vehicles in traffic (or traffic-like conditions), observe their performance and use these observations as the basis for a statistical comparison of the relative safety of human versus automated driving (Kalra and Paddock, 2016). This approach has its merits and certainly is a methodologically robust way of assessing safety performance (if biases are skilfully addressed as discussed later) but the sheer volume of kilometres driven necessary to infer statistically correct findings. According to (Kalra and Paddock, 2016), automated driving systems would have to log hundreds of millions –potentially hundreds of billions – of miles in order to demonstrate their innocuity in terms of fatalities and severe injuries. Under even aggressive testing assumptions, this would take tens –possibly hundreds – of years to accumulate the necessary evidence. It seems clear that technology developers (and licensing authorities) cannot simply "drive their way" to safety (Kalra and Paddock, 2016) (Rand 2016) and that alternative safety-validation frameworks are needed.

One approach is to pool safety-performance related test vehicle data together across all operators. Currently this data is closely guarded by industry, as it is seen as highly-sensitive commercial information and ultimately forms the basis for their competitive edge in the market. While commercial interests need to be protected, a way needs to be found to open up data to authorities and regulators to guarantee maximum road safety levels. There are emerging models that enable analytical processing of data without revealing commercially sensitive or privacy revealing data – e.g. by leveraging third-party vetting or developing distributed ledger technology-based approaches like Directed-Acyclic-Graph Technologies (DAG) (ITF, 2018). DAG scales well, can quickly validate transactions, is designed to be deployed in machine-to-

machine applications and can run on sensors and cyber physical systems. These types of approaches may enable a more data-driven governance model, offering the flexibility required in this fast-moving field.

A second strategy is to build up driving "experience" in virtual or simulator-based environments. This approach is typically used in system development to rapidly acquire knowledge and experience of a large variety of road environments and to develop greater reliability in system operations. Virtual kilometres, if they incorporate some form of randomisation, can also serve to "train" AI algorithms. As noted earlier, these approaches increase the volume of testing data without necessarily enriching its diversity (beyond what is programmed). As a result, simulator-generated test kilometres would be difficult to compare to actual test kilometres and safety-related assessment of automated driving systems would potentially be incomplete or biased. For the most part, those looking at testing are thinking mainly in terms of complex test tracks such as Mcity in Ann Arbor or the new high-speed track at Willow Run, Michigan (American Center for Mobility).

A final strategy, and one that seems to currently be utilised in certain testing environments as in several US States, is one that specifies the condition of automated system and behavioural outcomes rather than safety outcomes directly. In these cases, manufacturers must certify that their systems are safe and that reasonable and demonstrable precautions have been undertaken to make this so. This approach assumes a lot (perhaps too much in light of a recent fatal crash in Arizona) but could be more formalised and standardised than it is today One set of required system characteristics could include the following. Automated driving systems must (Reschka, 2016):

- Have condition awareness (both of their state and their surrounding context).
- Be aware of their current performance capabilities.
- Be aware of their current functional limitations in relation to the current situation.
- Be operated in a condition that reduces risk to passengers and other road users to an acceptable level.
- Be able to safely return to a normal operating state or gracefully degrade to a safe condition.

Statements of "acceptable" but non-quantified levels of risk necessarily involve ethical and political judgement. Jurisdictions promulgating condition and system behaviour rules for licensing automated driving tests should ensure consistency with overall road safety goals and those embodied in the testing requirements.

Finally, the issue of appropriate benchmarks is an important one to consider when evaluating the safety performance of automated driving systems. One clear issue today is that kilometres driven by automated driving systems are not representative of "average" driving kilometres in any other conditions than those faced in the testing environment. Concretely, if the majority of test kilometres are driven in generally sunny, clear, dry conditions on wide and relatively uncomplicated road networks and few complex interactions, the resulting safety performance should not be compared to average conventional driving which takes place in a wider range of contexts, conditions and levels of complexities. Clearly automated driving systems would be challenged by "average" driving conditions faced by all drivers and in all contexts and would not be able to function in many extreme conditions routinely faced by human drivers which is why testing environments are carefully selected. This is a sensible approach for testing but results in information that could bias safety assessments.

Likewise comparing automated driving performance to "average" human drivers (who are generally relatively safe) versus "risky" drivers may lead to different safety assessments. There is also an irony in that while the introduction of automated driving may improve overall safety by eliminating dangerous behaviour by a small set of unskilled, risky or impaired drivers, it may increase relative risk faced by

passengers of automated vehicles who previously displayed good driving behaviour on average (Noy et al., 2017; Kalra and Paddock, 2016). This may impact the uptake of automated driving. As safety-related driver behaviour is improving in many countries over time, the appropriate temporal benchmark for safety assessments of the future roll-out of automated driving systems come into play. Should driving system performance of these systems be compared to current human drivers or likely safer future human drivers? This too will have an incidence on the assessment of the overall road traffic system over time.

Difficulty in regulating a fast-moving target

Whereas there is broad precedent for, and experience with, validating the safety performance of hardware systems for cars and trucks, and hardware-software systems for aircraft, trains and other rail-based vehicles such a subway and light-rail systems, this is not the case of automated driving systems whose performance is equally linked to hardware elements (sensors, processors, actuators and traditional vehicle control mechanisms and surfaces) as it is to software. Indeed, for any given automated vehicle, regulatory oversight is fully present for much of the analogue systems present (chassis, protective equipment, steering and acceleration/braking control, body geometry and composition, etc.) and wholly absent for the lines of code that control the operation of the vehicle in some, or all contexts.

For traditional driving, the driver must be trained and licensed with increasing levels of training for increasingly complex and/or dangerous vehicles like trucks. There is no analogue for licensing software that controls vehicle operation – even for critical driving tasks like steering and acceleration/deceleration. Part of the reason for this is that the emergence of automated driving systems is quite recent and there is no consensus on the best regulatory model to apply for licensing these systems.

A second, more challenging issue is that the way in which the automated driving system operates is inextricably linked to its algorithmic skill. Given how easy it is to modify and update performance- and safety-critical code that can radically change the operating parameters and performance of automated driving systems, there is little certainty as to what is the precise object of regulation. A licensed automated driving system may have one set of acceptable (or expected) performance characteristics whereas the same hardware/software system may display completely different performance characteristics following an update of its code. Unless the entire system is re-validated and re-licensed, the system operating on the road is no longer that which was originally licensed. This raises the issue of how well regulators retain control over desired system characteristics with regards to safety. It also raises the issue of how well (human) drivers and passengers, other automated driving systems and, ultimately, regulators themselves share a common expectation and understanding of the performance and driving characteristics of constantly changing automated driving systems.

From the Safe System perspective, there is an imperative to ensure high levels of *expectancy* in the driving environment. While it may certainly prove challenging to require standard coding for some aspects of automated driving system operations, it may prove worthwhile standardising some aspects of automated driving behaviours (e.g. rules relating to right-of-way, acceleration and deceleration rates in normal operation, signalling intent, etc.) that increase the overall predictability of the driving environment – including for non-automated traffic participants (cars, trucks motorcyclists and cyclists).

How much safety is enough?

Much of the discourse around automated driving points to how bad human-based driving performance is today. This raises the question of how much safety is safe enough when compared to today's traffic environment? This question is probably one of the most challenging when discussing the introduction of automated vehicles. Discussion of this issue unavoidably will have an ethical dimension.

From the perspective of both society and individuals, there seems to be a high value placed on protecting *passengers* than *drivers*. Being an innocent victim is significantly different from being an active agent, a

driver. The effect of this is that safety in trains and planes is significantly higher than in cars. In aviation and for trains there is virtually no balancing between safety and efficiency. Safety comes first. In the road transport system such balancing is still common practice even if Vision Zero is slowly changing practice. The Safe System approach substitutes such arbitrary trade-offs with choices as to how to reduce casualties. The clearest example is the choice to either reduce speed or invest in upgrading infrastructure to be safe at higher speeds.

For automated cars it seems relevant to put the safety ambitions as high as the levels for aviation or rail travel, at a twentieth of the risk of today's car riding. At a very minimum, the "target" safety level for automated cars, should be at least as good as traffic in the safest countries and, as this is an evolving target (the safety package of most modern cars is significant and is continually improving), so too should the safety performance of automated driving.

The road transport system today runs with drivers not fully adhering to the rules. But taking the rules literary will probably be a prerequisite for automated vehicles. The Vienna convention of road traffic from 1968 sets the framework for road regulation in most countries. One significant article in the convention in the context of safety and security of automated vehicles is shown in Box 4.

Box 4. **Excerpt from 1968 Convention on Road Traffic**

Article 13

"Every driver of a vehicle shall in all circumstances have his vehicle under control so as to be able to exercise due and proper care and to be at all times in a position to perform all manoeuvres required of him. He shall, when adjusting the speed of his vehicle, pay constant regard to the circumstances, in particular the lie of the land, the state of the road, the condition and load of his vehicle, the weather conditions and the density of traffic, so as to be able to stop his vehicle within his range of forward vision and short of any foreseeable obstruction. He shall slow down and if necessary stop whenever circumstances so require, and particularly when visibility is not good."

Source: UNECE Sustainable Transport Division

The key aspect is that the driver (or in the case of automated cars the control mechanism) always should adapt vehicular speed so to be able to stop for any foreseeable obstruction. This article should be implemented in all national regulation in the contracting countries as it simply adapts what is stipulated for human drivers to automated driving systems. The RSS framework described earlier and in Box 3 is one way of formalising this imperative.

Human drivers contravene the requirement that they be able to stop within the range of forward vision. This is clearly the case in the dark, rainy or foggy traffic situations. Human drivers also undertake risky behaviour such as overtaking buses at bus stops even though there is a risk of sudden pedestrian ingress into the roadway. Humans often take risks that automated vehicles in most cases are incapable of taking (unless programmed to do so).

Combining the demands from the Vienna convention and the imperatives of the Safe System approach helps to guide operational parameters for automated driving systems. In order to meet these imperatives, the automated driving system must always plan and act so as to remain within the normal driving envelope. The energy level can never be higher than the allowed speed limit, but it is further restricted by the demand to be able to stop short of any foreseeable obstruction. The sensors and their limitations will restrict possible speeds. One must keep in mind that the Vienna convention demands a crash free system, not an injury free traffic. The automated vehicle will therefore likely move slower than the rest of the traffic, especially in the presence of pedestrians and cyclists and it may be sometimes a better idea to exclude human-driven cars from some environments.

Another possibility would be to dedicate restricted driving environments for automated vehicles. This serves to separate potential crash opponents but the provision of separate, dedicated infrastructure in many urban

contexts runs the risk of creating new barriers to the use urban space, assigning even more space to vehicles at the expense of pedestrians, cyclists and using space for non-transport purposes. The US National Association of City Transport Officials (NACTO, 2017) sets out strategies to adapt automated vehicle use to cities and not the other way around.

Even if embedded kinetic energy levels are within system tolerances, automated vehicles will still be sharing the road with other vehicles and road users. Crashes will occur and thus automated vehicles need, at least for now, all of the embarked safety technologies required for "normal" cars and trucks. It is also likely that the public will have greater difficulty in accepting fatalities and severe injuries resulting from automated vehicle versus human-driven vehicle crashes. As societal demands for safety increase, the most severe injury that is acceptable will be at lower bounds of those experienced today.

In the hopefully few and rare crashes that automated vehicles experience, injury levels must be low. Infrequent and innocuous crashes will likely prove to be a prerequisite for public acceptance. These vehicles must act and feel like reliable and trustful traffic elements. Demands regarding the safety performance of the road transport system may also increase over time such that, for example, even children may be able to walk or cycle without risking any injury as a result of a conflict with automated vehicles. These and similar demands will likely condition society's tolerance of automated driving behaviours.

Box 5. **National Association of City Transportation Officials Blueprint for Autonomous Urbanism**

The US-based National Association of City Transportation Officials has issued guidance to help ensure that city officials and industry work towards a shared vision of the role that automation can play in making cities more liveable, efficient and sustainable. The aspirational blueprint is articulated around 6 principles:

Safety is the top priority

Streets should be designed for safety of all users, with special attention needed for pedestrians and cyclists. Cities should require highly automated vehicles be programmed for safe, slow speeds on streets, with mandatory yielding to people outside of vehicles. Maximum vehicle operating speeds in city street environments should not exceed 20mph, or 25mph in limited circumstances, with lower speeds in downtown and neighbourhood zones.

Provide mobility for the whole city

The benefits of autonomous urbanism can only be realized if mobility is made more accessible, convenient, and affordable for the entire city. Cities and their partners should offer flexible and affordable mobility options tailored to the needs of different communities, from walking and biking to fixed transit and ridesharing.

Rebalance the right-of-way

With the right policies, autonomous vehicles can move more people in fewer vehicles on less congested streets. That means that cities can use space more wisely. Instead of planning for roadway expansion, reallocate street space to active, sustainable modes and use technology to manage the public realm dynamically.

Manage streets in real-time

New technology makes real-time, proactive street management feasible. Cities must leverage this opportunity to revolutionize the services they provide and the ways they capture revenue. Real-time right-of-way management and vehicle occupancy pricing mechanisms will allow cities to incentivize shared and active modes over private automobile trips, while reapportioning vehicle space as public space.

Move more with fewer vehicles

As technology is embedded in urban transportation, vehicles can assume maximum rider occupancy, creating an interconnected network of mobility supply and demand for freight or passengers. Transit agencies will need to adapt to new consumer expectations and reshape services to ensure seamless connections with other modes.

Public benefit guides private action

Autonomous urbanism should foster balanced collaboration with the private sector that maximizes public benefit. Smart governance ensures that these partnerships are neither unconditional endorsements nor punitive prohibitions, but are instead guided by set criteria and clear, measurable and adaptive policy goals.

Adapted from: (NACTO 2017)

Safety implications of cybersecurity-related threats to automated driving

As early as 2016 the cybersecurity vulnerabilities of connected cars were revealed through a series of highly publicised cyber-attacks against a commercially available vehicle in the United States (IOActive, 2015; Greenberg, 2016). In these, the vehicle's safety-critical performance was compromised rendering the driver completely defenceless against the cyber-attack and the vehicle unable to manoeuvre safely in heavy traffic. Other car-related cybersecurity vulnerabilities had already been demonstrated, but not for critical sub-systems.

In 2015, a German premium car brand suffered a cyber-attack that exploited the connectivity offered to car owners to remotely unlock or communicate with the car. The communication protocols were easy to compromise – despite requiring rigorous identity authentication the manufacturer failed to protect the identity of the vehicle and failed to securely encrypt the communication between the unlocking mobile device and the car (Spaar, 2015). Researchers demonstrated that with minimum effort and hardware resources, they were able to intercept the messages between the endpoints and to spoof requests for opening or unlocking the car within minutes. In this case, the manufacturer had assigned the same cryptographic key to identify and authenticate endpoints, rather than allocate unique cryptographic keys for all participants (including owners and their cars). As a result of the hack, the vendor patched the basic shortcomings but has yet to prove that sustainable measures have been established to mitigate the risk of abusing communications.

Only a few years ago the vehicles on streets were mostly fully independent and disconnected machines that were operated by a human being using his senses and controlling how the vehicle moved by simple interactions. Then access to the cars' electronic control units (ECUs) was standardised by introducing the OBD-2 (OBD: on-board diagnostics) interface and the first malevolent abuse cases surfaced. These resulted in the unauthorised access to previously protected information stored in cars' proprietary control units. The good intent of providing such a standard connector to enable third parties to be able to service vehicles without recourse to prohibitively costly special equipment was abused due to missing security controls in the design.

A growing number of hacking tools and manipulation software has become available, enabling direct interaction with ECUs – with potentially severe impacts for car safety systems and for drivers. Fortunately, the OBD-2 interface is usually located well out of reach of casual hackers, inside the dashboard of a car. But the advent of cheap wireless adaptors for OBD-2 has now created a more critical attack surface for hackers: either local Bluetooth or even far reaching mobile network connectivity is offered at very low cost, opening large unprotected gateways into the inner circuitry of the car, if not configured in a secure manner.

Besides these simple, user-initiated weaknesses, a modern car features a plethora of possible attack vectors presented by the sheer number of wireless and wired connections it offers: on-board Wi-Fi, Bluetooth to connect mobile devices, SD card readers, GPS sensors, radar/ lidar sensors, ultrasonic sensors, 3G/ LTE connectivity and even CD/ DVD players offer a wide range of possible methods to tamper with the security and integrity of a modern vehicle. Widespread "featurism" that promotes new functions for connected cars also increases the risk of adding unsecured, unstable and thus critically dangerous code to ECUs inside the car.

Over recent years, the cybersecurity threat landscape has broadened considerably, with availability of tools and techniques for free download, sale or rental through the "dark web" and other brokerage and exchange

sites. In particular, criminal organisations have begun to provide both retail and bespoke services granting wider access to sophisticated tools.

One thing that seems certain going forward is that the safety performance of the road traffic system in the presence of highly automated vehicles will no longer solely depend on the combined safety performance of component hardware and software elements. It will also depend on how robust automated driving systems are to malevolent attacks – especially when the design of automated driving systems may create new systemic cybersecurity-related vulnerabilities.

There is no single cybersecurity threat that may target automated driving systems or the traffic system more generally. Potential threats are multiple, motivations varied and capabilities uneven. Threats arise from state and quasi-state actors, single-issue activists, so called "hobbyists", and (organised) criminals. Not all of these actors would necessarily seek to directly or indirectly provoke a crash (or wide-scale system failures that might entail crashes), but some would. If vulnerabilities are left unaddressed, such attacks could be carried out resulting in negative safety outcomes.

Key attack vectors, or vulnerabilities, range from upstream designers, manufacturers and vendors all the way to the end-points – the vehicles themselves - comprised of hardware and, especially, software sub-systems. These threats are common to many forms of automated driving irrespective of particular use cases (logistics, public transport, shared mobility/other ride services or privately-owned vehicle usage). Emergent threats, vulnerabilities and consequent risks associated with the uptake and deployment of automated vehicles include:

- Designer vulnerability: Source code, architecture, component specification, and product whole life design and support.
- Manufacturer vulnerability: Component selection and manufacture (cheap/ potentially compromised), threat identification and mitigation, software/ firmware update creation, and version control.
- Vendor vulnerability: Inventory management, inventory protection, version management. A special consideration is the extent to which sensing and other critical sub-components are designed manufactured and programmed with attention to security.
- Maintainer vulnerability: Version management, design integrity management, platform protection, 3rd Party Engineering/Customisation/Enhancement Compatibility and Vulnerability Management.
- Infrastructure Provider Vulnerability: Direct network attack, jamming of communications and location services, spoofing, impersonation, and interfaces to/ from other public systems.
- Law enforcement and traffic management vulnerability: Direct network attack, jamming of communications and location services, spoofing, and impersonation.
- End point vulnerability: On-board interface (external or internal attack), individual vehicle, control, access, disruption of operation, selective/ non-selective, and ransom, kidnapping, or theft of data.

Comprehensive cybersecurity frameworks for automated driving

These multiple points of vulnerability underscore that complex "systems of systems", like those delivering automated driving, require comprehensive frameworks to ensure systemic cybersecurity. In 2017, the UK Department for Transport (DfT) in conjunction with the UK Centre for the Protection of National Infrastructure (CPNI) released such high-level guidance for the automotive sector, the automated driving and intelligent transportation system ecosystem and their collective suppliers (DfT, 2017). The "Key

Principles of Cyber Security for Connected and Automated Vehicles" outlines 8 fundamental building blocks that should underpin systemic cybersecurity best practice (Table 5).

These principles set out a comprehensive framework for addressing cybersecurity in the automated driving ecosystem but standards are required to deliver effective cybersecurity. SAE guidance J3061 (Cybersecurity guidebook for cyber-physical vehicle systems) and J3101 (Requirements for hardware protected security for ground vehicle applications), along with numerous ISO standards relating to identity management, authentication, securing information technology systems and privacy all form the base on which to build the operational framework for securing automated driving systems. The US Department of Transport's National Highway Traffic Safety Administration has also issued guidance on cybersecurity best practices for vehicles which builds on SAE and other recommendations (NHTSA 2016).

At the outset, however, two fundamental design strategies condition automated driving cybersecurity. These relate to the functional *isolation* or not of safety-critical subsystems and whether safe system performance is *conditioned on connectivity* to external networks. These are not trivial design decisions. The choice of strategy will have an incidence on whether imperatives for safety and cybersecurity can be reconciled – and if so, how easily or not.

Table 5: **Key Principles of Cyber Security for Connected and Automated Vehicles (United Kingdom)**

Principle 1 - organisational security is owned, governed and promoted at board level.
1.1: There is a security program which is aligned with an organisation's broader mission and objectives.
1.2: Personal accountability is held at the board level for product and system security (physical, personnel and cyber) and delegated appropriately and clearly throughout the organisation.
1.3: Awareness and training is implemented to embed a 'culture of security' to ensure individuals understand their role and responsibility in ITS/CAV system security.
1.4 All new designs embrace security by design. Secure design principles are followed in developing a secure ITS/CAV system, and all aspects of security (physical, personnel and cyber) are integrated into the product and service development process.
Principle 2 - security risks are assessed and managed appropriately and proportionately, including those specific to the supply chain.
2.1: Organisations must require knowledge and understanding of current and relevant threats and the engineering practices to mitigate them in their engineering roles.
2.2: Organisations collaborate and engage with appropriate third parties to enhance threat awareness and appropriate response planning.
2.3: Security risk assessment and management procedures are in place within the organisation. Appropriate processes for identification, categorisation, prioritisation, and treatment of security risks, including those from cyber, are developed.
2.4: Security risks specific to, and/or encompassing, supply chains, sub-contractors and service providers are identified and managed through design, specification and procurement practices.
Principle 3 - organisations need product aftercare and incident response to ensure systems are secure over their lifetime.
3.1: Organisations plan for how to maintain security over the lifetime of their systems, including any necessary after-sales support services.
3.2: Incident response plans are in place. Organisations plan for how to respond to potential compromise of safety critical assets, non-safety critical assets, and system malfunctions, and how to return affected systems to a safe and secure state.
3.3: There is an active programme in place to identify critical vulnerabilities and appropriate systems in place to mitigate them in a proportionate manner.
3.4: Organisations ensure their systems are able to support data forensics and the recovery of forensically robust, uniquely identifiable data. This may be used to identify the cause of any cyber, or other, incident.
Principle 4 - all organisations, including sub-contractors, suppliers and potential 3rd parties, work together to enhance the security of the system.
4.1: Organisations, including suppliers and 3rd parties, must be able to provide assurance, such as independent validation or certification, of their security processes and products (physical, personnel and cyber).
4.2: It is possible to ascertain and validate the authenticity and origin of all supplies within the supply chain.
4.3: Organisations jointly plan for how systems will safely and securely interact with external devices, connections (including the ecosystem), services (including maintenance), operations or control centres. This may include agreeing standards and data requirements.
4.4: Organisations identify and manage external dependencies. Where the accuracy or availability of sensor or external data

is critical to automated functions, secondary measures must also be employed.
Principle 5 - systems are designed using a defence-in-depth approach.
5.1: The security of the system does not rely on single points of failure, security by obscuration or anything which cannot be readily changed, should it be compromised.
5.2: The security architecture applies defence-in-depth and segmented techniques, seeking to mitigate risks with complementary controls such as monitoring, alerting, segregation, reducing attack surfaces (such as open internet ports), trust layers / boundaries and other security protocols.
5.3: Design controls to mediate transactions across trust boundaries, must be in place throughout the system. These include the least access principle, one-way data controls, full disk encryption and minimising shared data storage.
5.4: Remote and back-end systems, including cloud-based servers, which might provide access to a system have appropriate levels of protection and monitoring in place to prevent unauthorised access.
Principle 6 - the security of all software is managed throughout its lifetime.
6.1: Organisations adopt secure coding practices to proportionately manage risks from known and unknown vulnerabilities in software, including existing code libraries. Systems to manage, audit and test code are in place.
6.2: It must be possible to ascertain the status of all software, firmware and their configuration, including the version, revision and configuration data of all software components.
6.3: It's possible to safely and securely update software and return it to a known good state if it becomes corrupt.
6.4: Software adopts open design practices and peer reviewed code is used where possible. Source code is able to be shared where appropriate.
Principle 7 - the storage and transmission of data is secure and can be controlled.
7.1: Data must be sufficiently secure (confidentiality and integrity) when stored and transmitted so that only the intended recipient or system functions are able to receive and / or access it. Incoming communications are treated as unsecure until validated.
7.2: Personally identifiable data must be managed appropriately. This includes: what is stored (both on and off the ITS / CAV system) what is transmitted how it is used the control the data owner has over these processes. Where possible, data that is sent to other systems is sanitised.
7.3: Users are able to delete sensitive data held on systems and connected systems.
Principle 8 - the system is designed to be resilient to attacks and respond appropriately when its defences or sensors fail.
8.1: The system must be able to withstand receiving corrupt, invalid or malicious data or commands via its external and internal interfaces while remaining available for primary use. This includes sensor jamming or spoofing.
8.2: Systems are resilient and fail-safe if safety-critical functions are compromised or cease to work. The mechanism is proportionate to the risk. The systems are able to respond appropriately if non-safety critical functions fail.

Source: (UK DfT, 2017)

Need for critical sub-system isolation

The control functions of an automated driving system rely on a complex and highly integrated network of dozens of sensors, actuators and microcontrollers. Besides creating issues of reliability and redundancy as a whole, each and every ingredient of this system also is a potential entry point for cyber-attacks. Such attacks could consist of manipulating controller protocols, modifying power circuits, changing data connections, or even impairing complete devices.

Consequently, cybersecurity does not only mean protecting data communication emanating to and from vehicles, but it also has to prevent unauthorised access to individual devices and microcontrollers or access to networks of such components in the vehicle.

In this respect, the discussion surrounding the cybersecurity vulnerabilities of automated driving systems is not dissimilar to discussions surrounding the security (and cybersecurity) of other complex systems within and outside of the transport sector (aircraft, train and metro systems, nuclear power plants, etc.) (Le Lann, 2017). In all of these systems, core safety-critical components are *isolated* on both a hardware and software level from non-critical components. In most cases, redundancies are built in to ensure critical sub-system performance even in degraded conditions.

In practical terms, automated driving safety-critical subsystems including steering control, acceleration and deceleration, should be isolated from others with independent processors, system memory, system architecture and separate (and redundant) power supply. The operating system governing these functions should undergo specific and robust cybersecurity vetting. Secure protocols are necessary for handling update policies for these systems (updates which should be the exception, rather than the rule). One part of the vetting should be to assess the cybersecurity risks of open-source code that is often bundled into various control and operating system software. Safety-critical subsystems should also integrate tamper-proof devices with independent state awareness to give the alert if the case of malevolent or accidental access to critical systems (Le Lann, 2017), (Le Lann 2018).

There is little formal agreement today as to what constitutes safety-critical subsystems but this is one area where accelerated work in the appropriate standard-setting bodies can prove helpful.

Autonomous or connected: What is safest?

As noted in the previous section, there are safety improvements to be gained from extending sensor fusion strategies to encompass vehicle-to-vehicle and vehicle-to-infrastructure connectivity. Some argue that for SAE level 4 and 5 automation, communication is not just an enabling feature but a necessary component of the vehicle control system, particularly in highly complex urban traffic situations. This line of thought has led the US National Transport Safety Board (NTSB) to the conclusion that automated vehicles require connectivity. Analysing a first fatal accident involving an automated vehicle the NTSB recommends that vehicles need to communicate with each other and that such a technology ought to be standardised and in every vehicle (NTSB, 2016). The Road Safety Council (DVR), a not-for-profit organisation and member of the European Road Safety Council (ETSC), also argues that vehicles should become cooperative and have the capacity to warn each other through direct and instant communication (DVR, 2017). The importance of connectivity for automated driving is also at the core of the work of the Cooperative Intelligent Transport Systems Platform in Europe (C-ITS) (C-ITS, 2016)(C-ITS, 2017). The C-ITS vision fully leverages vehicle-to-vehicle, vehicle-to-infrastructure and vehicle-to-x connectivity to ensure useful and useable data exchange between all components of the road traffic system. For this to happen, the C-ITS platform considers that automated vehicles must be cooperative and connected vehicles and outlines challenges that must be overcome to realise this vision:

> *"The cooperative and connected elements will allow vehicles to receive, in real-time, in addition to the digital knowledge of the infrastructure already available in the vehicle (e.g. digital maps), key attributes of roads relevant for automated driving, with the aim of adding predictability on what to expect on the road ahead and enlarging the decision base for using automatic mode. The cooperative element is required, amongst others, to handle complex traffic situations. To go beyond awareness … a new set of technology agnostic C-ITS messages for collective perception needs to be standardised. This means that future vehicles will share what they see and all vehicles in range will see what they see collectively. To make this work a common operational environment for sharing such messages will need to be developed, including the context and the interpretation boundaries (such as the quality assumptions to quantify trustworthiness, precision, timeliness and reliability of information) for the receiving vehicle."* (C-ITS, 2017).

There clearly is a strong expectation that the safety of the road traffic system can be enhanced by connectivity between automated (indeed, all) vehicles. The safety benefits of connectivity include the ability to see beyond human line of sight, to be apprised of the reactions of other vehicles and traffic participants to unforeseen or emergency situations, to more accurately detect pedestrians and cyclists, and "understand" complex behaviour patterns on the basis of immediate trajectory histories. Connectivity also

provides access to data that can be used to validate vehicle sensor data and hence increase trust that the automated driving system is making a decision on the basis of correct contextual awareness.

Where this consensus begins to break down is when considering if safety of the (automated) driving task should be made *dependent* on connectivity. This is a particularly acute question for automated driving systems because connectivity raises critical questions about the ability of networked automated driving systems to withstand cyberattacks that could compromise safety.

Proponents of the "connected" view have pushed to develop standards for connectivity-related hardware and communication protocols – and in the case of the United States – to mandate their use. Among these is the *Wireless Access in Vehicular Environments* (WAVE) framework encompassing IEEE 802.11p (and its European equivalent ETSI ITS-G5) and IEEE 1609.4 (US DOT 2009) (Dimitrakopoulos 2017). NHTSA proposed in 2016 that, starting in 2020, all new cars sold in the United States should have an on-board hardware/software system capable of communicating using the WAVE standard (NHTSA, 2015). However, this proposal seems likely to be delayed and, possibly, not fulfilled (Beene 2017) (Lowy 2017).

Proponents of light, non-safety-vital connectivity, on the other hand, point to long-standing practice in managing risk in critical systems. A core design principle for these systems is that in no case should the avoidance of unwanted outcomes (crashes in the case of automated driving systems) *rely on access to shared external communication channels*. These systems are designed to operate and fail safely on their own. One of the reasons for this is that communication networks are not designed with safety-critical functioning in mind and may not prove sufficiently reliable in times of crises (Le Lann, 2017).

A secondary reason is that connected systems – especially those that provide direct or indirect access to safety-critical sub-systems – create new potential cyber-attack surfaces. A critical issue here is how to define the "trust" boundary of critical sub-systems (Macher, et al. 2017). This boundary may be evaluated using a static, layer-based approach that identifies where each critical boundary lays in the system architecture and developing a "defence-in-depth" response. Alternatively, critical functional boundaries can be identified looking at processes, not system architecture, using a static threat assessment approach. In either case, trust boundaries must be robust and their violation instantly perceivable. The trust boundary can conceivably extend beyond the critical sub-systems of automated driving systems (see discussion further on managing trust) but doing so runs the risk of creating new attack surfaces. These are manageable risks – but they remain risks nonetheless and it is uncertain how well they can be managed in large-scale fleets of automated vehicles as these have not yet been deployed (Macher, et al. 2017).

A fundamental vulnerability in trying to ensure the cybersecurity of connected and automated vehicle networks is that potential threat vectors are treated in much the same way as threats are treated for computer networks – completely in cyber- or virtual-code space. This overlooks the fact that networks of connected vehicles are in fact hardware/software communicative platforms that move in space (sometimes quite fast) and whose environment is constantly changing. The material reality of the vehicles' environment matters in terms of potential errors of communication, dropped messages or messages never received.

Because vehicles are in movement, low-latency in mission-critical communication is necessary but it is uncertain if their transmission can guarantee using existing approaches if the number of connected vehicles scales up dramatically. Another aspect of the "physical" reality of these systems is that non-complying or malevolent members of these networks must be safely excluded as soon as they are discovered. But there is no agreement on how this might happen and it is unclear if "as soon" is soon enough from a safety perspective - much damage can happen in a short amount of time (Le Lann, 2017), (Le Lann 2018). Whatever the strategy employed, it seems clear that regulation of these systems will have to be flexible enough to quickly leverage improved safety outcomes when they are demonstrated.

Flexible regulation reduces certainty on the return for companies' investments and this, in turn, may reduce innovation. At the same time, it opens up pathways for the deployment of new innovative approaches to

improve automated driving system cybersecurity and safety. Balancing between the two will involve re-assessing the sufficiency of current, technology-driven regulatory approaches to deliver safety, security and innovation. A shift to a safety outcome-based or performance-based approach may seem an attractive option and is in line with many other domains of governance (education, health, etc.). If this is the strategy employed, countries will have to consider what outcome they want the system to deliver – a reduction in deaths and injuries or zero road deaths and serious injuries. The pertinence of the Safe System approach in framing outcome-based automated driving safety and cybersecurity rules seems clear.

Managing trust

Irrespective of the level of connectivity that will be designed into automated driving systems -- from low and non-critical, to high and mission-critical -- robust identity management and authentication frameworks will need to perform up to the level required by the selected cybersecurity strategy.

A good starting point is to provide each entity with a secure, verifiable digital identity (ID). For vehicles, this could be analogous to a digital vehicle identification number (VIN). Deploying a common identity layer and identity authentication protocol that enables trustworthy interoperability and secure connectivity between all entities helps set the groundwork for different levels of connectivity that are "safe enough" or "secure enough" for specific use cases and missions.

For this to happen, stakeholders engaged with developing systems, components, infrastructure or complete autonomous vehicles must be participate in both standardisation activity as well as interoperability testing in order to make sure that the common basic level of trust can be established for each and every device taking part in autonomous vehicle communication. This includes OEMs that will deliver the autonomous vehicles to the customers. Vehicles need not only have a secure verifiable identity, but must also provide provable cyber security vetting and certification. Currently, this is far from being the norm and cost pressure on OEMs may lead some to neglect opportunities to improve cybersecurity by specifying certain security-improving hardware/software devices.

If OEMs are not sufficiently motivated to invest in cybersecurity protection (versus, for example, infotainment and consumer connectivity) then it seems challenging to motivate the inventors and developers of the necessary connected car infrastructure to invest in security mechanisms themselves.

Nevertheless, this second group of organisations is pivotal in creating a trustworthy landscape that enables autonomous vehicles to securely gather input from surrounding sensors, infrastructure components and potentially, traffic guidance systems. Secure trustworthy digital identities will help to forge such a landscape – especially where connectivity is designed into the automated driving system.

Certifying trust in connected environments

Proponents of the connected automated vehicle ecosystem in Europe have call for connectivity to be implemented for Cooperative Intelligent Transport System(s) (C-ITS). C-ITS combines elements of vehicle-to-vehicle communications (V2V), vehicle-to-infrastructure communications (V2I) and is based on the leveraging devices that are either built into the vehicle, built into portable devices or built into roadside infrastructure (e.g. C-ITS *stations*). These devices must communicate in a secure way and provide a wide range of different standardised services that require robust authenticity and integrity-checking.

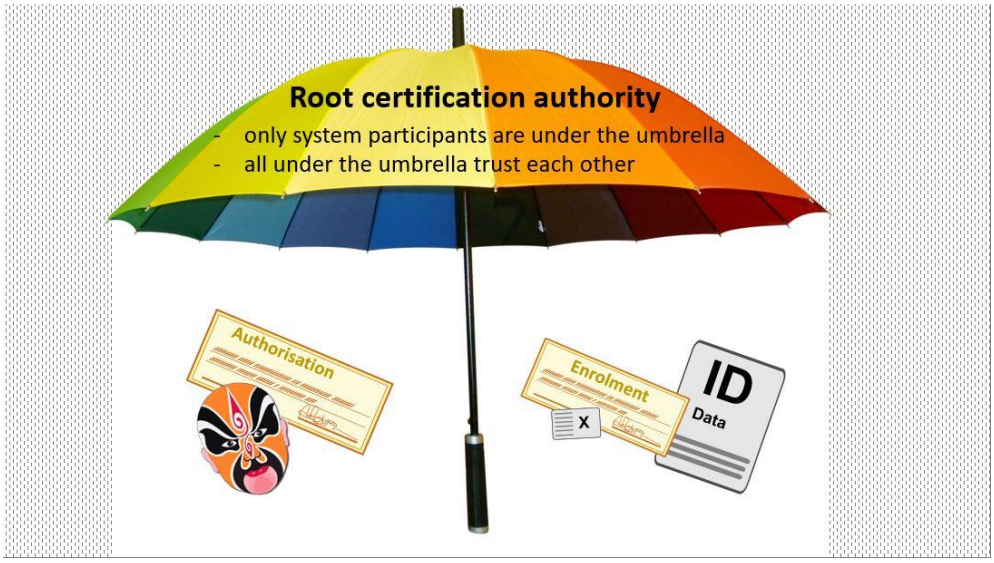
For this reason C-ITS messages are standardised and digitally signed so the receiver can verify the content and the integrity of the message before actually processing it. Only messages that comply with all agreed definitions and come from an authorised sender are processed. The authenticity of the sender is guaranteed by the digital certificate that is to be used to verify the signature. The digital certificate in turn needs to be trusted. The figures below use the analogy of the certificate as a mask to describe the European Public Key Infrastructure (PKI) to illustrate the principles behind the use of trusted certificates. In the European context, PKI also serves as a privacy-by-design measure which addresses privacy, as well as security.

Figure 3. **The philosophy of building and recognising trust**

Source: Kapsch TraffiCom

A common way to build trust between entities exchanging data is to define a trust model and implement it using PKI. A PKI is a system for issuing digital certificates to trusted devices through a hierarchy of entities, the authorisation entity and the enrolment authorities respectively. The highest level entity is the root certification authority and is authorised by an institution or a set of institutions. The root certification authority issues certificates to the subordinate authorisation and enrolment authorities and these in turn issue certificates to trusted devices such as C-ITS stations and devices. The use of PKI ensures that unauthorised participants can be immediately identified and that authorised participants who do not follow the rules can be expelled from the system.

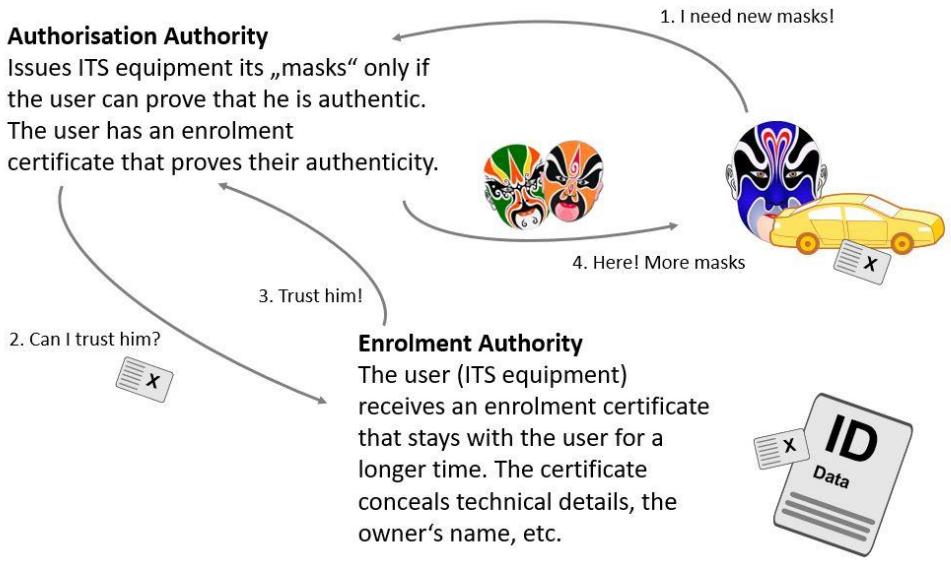
Figure 4.**The hierarchy of trust in a public infrastructure**



Source: Kapsch TraffiCom

The European Commission's C-ITS Platform has defined a trust model and certificate policy for the corresponding PKI for C-ITS in Europe (European Commission, 2017). This model ensures that participating entities do not reveal the identity of vehicles or vehicle owners when issuing certificates to vehicle-embedded ITS-stations. It also allows for frequent enough changes to the certificates presented to receivers such that tracking mobile ITS-stations (vehicles or portable devices) is thwarted (but not eliminated (Wiedersheim, et al. 2010)) thus helping to meet the requirements regarding unwanted revealing of personal data according to the EU General Data Protection Regulation (Figure 5).

Figure 5. **Changing certificates without revealing the identity of the motorist**



Source: Kapsch TraffiCom

Intelligent Transport applications have very high requirements in terms of high amount of data to be processed and low latency in processing. For this reason IEEE has published a standard that defines the formats and processes to build and verify signature and certificates that are specifically suitable for implementation on high performance and mobile embedded platforms. This standard - *Wireless Access in Vehicular Environment - Security Services for Applications and Management Messages* - IEEE 1609.2 - allows the processing of messages from hundreds or even thousands of vehicle per second. IEEE 1609.2 has been adopted by the European Commission's C-ITS Platform through its sister standard, ETSI TS 103 097, which is fully based on IEEE 1609.2 (European Commission 2017). These standards define secure data formats that reduce the overhead in traditional (X.509) certificates and rely on the well-known elliptic curve cryptography to speed up computations. The use of standardised and secured messages and the trust built up by a PKI operated according to a common policy promises to allow automated vehicles to exchange data securely between them, with the infrastructure and with other road participants.

Some uncertainty persists with regards to the adequacy of current certificate-based systems to effectively manage *safety-critical* data – especially for use cases that are built on periodic "beaconing" of safety-critical messages. If each beacon broadcast must be signed to be considered legitimate, valid certificates are used up at a very high rate putting pressure to provide them in sufficient number and at the time when they are needed (Petit, et al. 2015) (Le Lann 2018). Furthermore, even in the case of prevailing standards today, the validation of such certificates is "computationally expensive", which means even advanced computer systems need a few fractions of a second to test the trustworthiness of the certificate. In a world of scaled up and ubiquitous automated driving these fractions of a second are "ages" - as decisions need to be

computed in milliseconds and address the uncertainty of the vehicle's ever-changing environment as it moves.

Standards may evolve or new standards and approaches to authenticating the validity of safety critical messages may be deployed. For instance, there are new emerging approaches to providing trustworthy identity authentication for mobile connected autonomous vehicles, including certain forms of distributed ledger technologies that are purposely designed for managing networks of connected objects in close to real-time – such as Directed Acyclical Graphs (DAGs) (ITF, 2018).

Until these technologies are demonstrably ready to reliably handle high-volume and high-speed interactions in line with safety objectives, the Safe System approach would imply falling back on proven approaches or to ensure that essential safety performance is not predicated on connectivity.

Bibliography

Anderson, J.M., Kalra, N., Stanley, K.D., Sorensen, P., Samaras, C., Oluwatola, T.A. (2016), "Autonomous Vehicle Technology: A Guide for Policymakers", RAND Corporation.

Bainbridge, L. (1983), "Ironies of automation." *Automatica* 19, no. 6.

Beene, R. (2017), *Federal V2V mandate meets growing resistance,* http://www.autonews.com/article/20170417/OEM06/170419865/?templ (accessed May 7, 2018).

Blanco, M., Atwood, J., Russell, S., Trimble, T., McClafferty, J., Perez, M. (2016), "Automated vehicle crash rate comparison using naturalistic data", Final Report of the Virginia Tech Transportation Institute, https://vtechworks.lib.vt.edu/bitstream/handle/10919/64420/Automated%20Vehicle%20Crash%20Rate%20Comparison%20Using%20Naturalistic%20Data_Final%20Report_20160107.pdf?sequence=1&isAllowed=y (accessed 8 May, 2018)

C-ITS Platform, Cooperative Intelligent Transport Systems towards Cooperative (2016), Connected and Automated Mobility, Phase I, Final Report, European Commission, Brussels.

C-ITS Platform, Cooperative Intelligent Transport Systems towards Cooperative (2017), Connected and Automated Mobility, Phase II, Final Report, European Commission, Brussels.

Cummings, (2014) M.,"Man versus Machine or Man + Machine?", Intelligent Systems, IEEE. 29.

de Winter,J.C., Dodou, D. (2014), "Why the Fitts list has persisted throughout the history of function allocation" Cogn. Technol. Work, 16.

DfT (2017), The Key Principles of Cyber Security for Connected and Automated Vehicles, UK Department for Transport, UK Centre for the Protection of National Infrastructure, https://www.gov.uk/government/publications/principles-of-cyber-security-for-connected-and-automated-vehicles/the-key-principles-of-vehicle-cyber-security-for-connected-and-automated-vehicles (accessed May, 8, 2018)

Dimitrakopoulos, G. (2017), *Current Technologies in Vehicular Communication: Vehicular Communications Standards.* Springer International Publishing AG.

Dingus, T.A., Guo, F., Lee, S., Antin,J.F., Perez, M., Buchanan-King, M., and Hankey, J. (2016), "Driver crash risk factors and prevalence evaluation using naturalistic driving data", Proc. Natl. Acad. Sci., 113.

Dreany, H.H., Roncace, R., Young, P. (2018), "Safety engineering of computational cognitive architectures within safety-critical systems", Safety Science, Volume 103.

Deutscher Verkehrssicherheistsrat (DVR) (2017), Erhöhung der Verkehrssicherheit durch Vehicle-2-X-Kommunikation, https://www.dvr.de/dvr/beschluesse/2017-erhoehung-der-verkehrssicherheit-durch-vehicle-2-x-kommunikation.html

Eckhoff, D., N. Sofra, and R. German (2013), "A Performance Study of Cooperative Awareness in ETSI ITS G5 and IEEE WAVE." 10th Annual Conference on Wireless On-Demand Network Systems and Services (WONS). IEEE.

European Commission (2017), "Certificate Policy for Deployment and Operation of European Cooperative Intelligent Transport Systems (C-ITS) - Release 1." Brussels: C-ITS Platform chaired by the European Commission, June.

Greenberg, A. (2016), "The Jeep hackers are back to prove car hacking can get much worse", WIRED article, https://www.wired.com/2016/08/jeep-hackers-return-high-speed-steering-acceleration-hacks/

Härri, J., and F. Berens (2017), "Challenges and Opportunities of WiFi-based V2X Communications." Berlin.

Hendricks, D.L., J.C. Fell, and M. Freedman (2001). The relative frequency of unsafe driving acts in serious injury accidents. Veridian Engineering, US Department of Transportation National Highway Transportation Safety Administration.

Fagnant,D.J., Kockelman, K. (2015), "Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations", Transport. Res. Part A: Policy and Practice, 77.

Fitts, P.M. (1951), "Human engineering for an effective air navigation and traffic control system", National Research Council, Washington, DC.

IOActive Security Advisory (2015), Harman-Kardon UConnect Vulnerability, https://ioactive.com/wp-content/uploads/2018/05/IOActive_Advisory_Harman-Kardon.pdf

ITF (2008), "Ambitious road safety targets and the safe system approach." International Transport Forum at the OECD, Paris.

ITF (2015), " Automated and Autonomous Driving: Regulation under uncertainty", International Transport Forum at the OECD, Paris.

ITF (2016), "Zero Road Deaths and Serious Injuries: Leading a Paradigm Shift to a Safe System." International Transport Forum at the OECD, Paris.

ITF (2018a), Speed and Crash Risk. IRTAD, International Transport Forum at the OECD, Paris.

ITF (2018b), "Blockchain and Beyond: Coding 21st Century Mobility", International Transport Forum at the OECD, Paris.

Kalra, N., Paddock, S.M. (2016), "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?", Transportation Research Part A: Policy and Practice, Volume 94.

Le Lann, G. (2017), Protection de la vie privée, innocuité et immunité envers les cybermenaces dans les futurs réseaux de véhicules autonomes connectés, C&ESAR 2017 - Protection des données face à la menace cyber.

Le Lann, G. (2018), "Autonomic Vehicular Networks: Safety, Privacy, Cybersecurity and Societal Issues". IEEE Vehicular Technology Conference Spring 2018 -- First International Workshop on research advances in Cooperative ITS cyber security and privacy (C-ITSec), Jun 2018, Porto, Portugal.

Lonsdale Systems (2016), "Software quality essentials", http://lonsdalesystems.com/site/course/software_quality_essentials.php (accessed 8 May, 2018)

Lowy, J. (1 November 2017), *APNewsBreak: Gov't Won't Pursue Talking Car Mandate.* https://www.usnews.com/news/business/articles/2017-11-01/ap-newsbreak-govt-wont-pursue-talking-car-mandate (accessed May 7, 2018).

Macher, G., R. Messnarz, E. Armengaud, A. Riel, E. Brenner, and C. Kreiner (2017). "Integrated safety and security developments in the automotive domain." *SAE Technical Paper*.

NACTO (2017), "Blueprint for Autonomous Urbanism." New York City: National Association of City Transport Officials.

NHTSA (2016), "Cybersecurity Best Practices for Modern Vehicles." Washington, DC: US Department of Transportation National Highway Traffic Safety Administration.

NTSB (2016), "Preliminary Report, Highway HWY16FH018", US National Traffic Safety Board, https://www.ntsb.gov/investigations/AccidentReports/Pages/HWY16FH018-preliminary.aspx (accessed May 8, 2018).

Ni, R., Leung, J. (2016) "Safety and Liability of autonomous vehicle technologies", MIT.

Noy, I.Y., D. Shinar, and W.J. Horrey (2018), "Automated driving: Safety blind spots." *Safety Science,* 102.

Otte, D., B. Pund, and M. Jansch (2009), "A new approach to accident causation analysis by seven steps." *ACASS 21st ESV Conference.* Stuttgart, Germany.

Parasuraman, R., Sheridan, T.B., Wickens, C.D. (2000), "A model for types and levels of human interaction with automation" , IEEE Trans. Syst., Man, and Cybernetics - Part A: Syst. Hum., 30.

Petit, J., F. Schaub, M. Feiri, and F. Kargl (2015), "Pseudonym schemes in vehicular networks: a survey." *IEEE Communication Surveys and Tutorials.* Vol. 17.

Rand (2016). "Driving to safety-How many miles of driving."

Reader, R. (2018), "Who's Making Sure That Self-Driving Cars Are Safe?", Fast Company, https://www.fastcompany.com/40548215/whos-making-sure-that-self-driving-cars-are-safe (accessed May 8, 2018).

Reschka, A.(2016), Safety Concept for Autonomous Vehicles. In: Maurer M., Gerdes J., Lenz B., Winner H. (eds) Autonomous Driving. Springer, Berlin, Heidelberg.

SAE International (30 September, 2016), "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles - J3016_201609", https://www.sae.org/standards/content/j3016_201609 (accessed May 5, 2018).

Sabey, B.E., and G.C. Staughton (1975). "Interacting roles of road, environment and road user in accidents." *Fifth international conference of the international association for accident and traffic medecine and the third international conference on drug abuse of the international council on alcohol and addiction.* London.

Shalev-Shwartz, S., S. Shammah, and S. Shashua (2018), "On a Formal Model of Safe and Scalable Self-driving Cars." Edited by Cornell University. arXiv.org Computer Science, Robotics.

Shinar, D. (2017), Traffic safety and human behavior, (2nd ed.), Emerald Publishing, Bingley, UK.

Shoettle, B., and M. Sivak (2015), "A preliminary analysis of real-world crashes involving self-driving vehicles.", UMTRI-2015-34, University of Michigan, Transportation Research Institute, Ann Arbor.

Shoettle, B. (2017), "Sensor fusion: A comparison of sensing capabilities of human drivers and highly automated vehicles", SWT-2017-12, University of Michigan, Transportation Research Institute, Ann Arbor.

Singh, S. (2015), "Critical reasons for crashes investigated in the national motor vehicle crash causation survey." *Traffic Safety Facts Crash Stats Fact Sheet.* Washington, D.C.: US Department of Transportation.

Spaar, D. (2015), "Auto, öffne dich!", CT Magazin für computer technik, https://www.heise.de/ct/ausgabe/2015-5-Sicherheitsluecken-bei-BMWs-ConnectedDrive-2536384.html

Stolte, T., Hosse, R.S., Becker, U., Maurer, M. (2016), "On Functional Safety of Vehicle Actuation Systems in the Context of Automated Driving", IFAC-PapersOnLine, Vol.49(11).

US DOT. (2009), *IEEE 1609 - Family of Standards for Wireless Access in Vehicular Environments (WAVE).* https://www.standards.its.dot.gov/Factsheets/Factsheet/80 (accessed May 5, 2018).

VDA (2015), "Automation: From automated driver assistance systems to automated driving." *VDA Magazine*.

Wegman, F., and L.T. Aarts (2006). *Advancing Sustainable Safety: National Road Safety Outlook for2005-2020.* Leidschendam: Dutch Institute of Road Safety Research (SWOV).

WHO (2015), "Global Status Report on Road Safety", United Nations World Health Organization, Geneva.

Wickens, C.D., Gordon-Becker, S., Liu, Y., Lee, J.D. (2003), An introduction to human factors engineering (second ed.), Pearson Prentice Hall, Upper Saddle River, NJ.

Wiedersheim, B., M. Ma, F. Kargl, and P. Papadimitratos (2010). "Privacy in inter-vehicular networks: why simple pseudonym change is not enough." Kranjska Gora, Slovenia: IEEE.

Wu, X., Subramanian , S., Guha, R., White, R.G., Li, J., Lu, K.W., Bucceri, A., Zhang, T, (2013), "Vehicular Communications Using DSRC: Challenges, Enhancements, and Evolution." *IEEE Journal on Selected Areas in Communications*.

International Transport Forum

Safer Roads with Automated Vehicles?

This report examines how increasing automation of cars and trucks could affect road safety, and which security vulnerabilities will need to be addressed with the rise of self-driving vehicles. It applies the principles of the Safe System approach and relevance of Vision Zero for road safety to the wider discussion on vehicle automation. It also takes into consideration the security of the cyber-physical system associated with automated driving, including a definition of relevant system boundaries and future-proof minimum requirements for safety and security.

The work for this report was carried out in the context of a project initiated and funded by the International Transport Forum's Corporate Partnership Board (CPB). CPB projects are designed to enrich policy discussion with a business perspective. Led by the ITF, work is carried out in a collaborative fashion in working groups consisting of CPB member companies, external experts and ITF researchers.

International Transport Forum
2 rue André Pascal
75775 Paris Cedex 16
France
T +33 (0)1 45 24 97 10
F +33 (0)1 45 24 13 22
Email : itf.contact@oecd.org
Web: www.internationaltransportforum.org