4rd International Conference on Public Policy (ICPP4)
June 26-28, 2019 – Montreal, Canada

Panel T13P04 Session 2

Governing AI and Autonomous Systems

Title of the paper

An examination of discrimination and safety and liability risks stemming from algorithmic decision-making in AVs

Authors

Hazel Lim and Araz Taeihagh

Lee Kuan Yew School of Public Policy, National University of Singapore

Friday, June 28th 2019 14:30 to 16:00 (Room: MB S2.455)

An examination of discrimination and safety and liability risks stemming from algorithmic decision-making in AVs

Hazel Lim and Araz Taeihagh[1]
Lee Kuan Yew School of Public Policy, National University of Singapore

Abstract – Algorithms are increasingly entrenched in how decisions are made in society today, particularly in autonomous systems – In transportation, algorithms drive the operation of autonomous vehicles (AVs) through its environmental perception, decision-making and decision execution to yield AVs' manifold societal benefits, particularly road safety. However, algorithmic decision-making is plagued by concerns over bias, ethical considerations, perverse incentives of stakeholders and technical issues that shape the programming and implementation of these algorithms. These issues stemming from algorithmic decision-making can translate into greater safety risks as well as potential societal discrimination imposed by AVs. Due to ambiguities in the delineation of responsibilities for the design, manufacture and operation of AVs, stakeholders in the supply chain are also exposed to new liability risks for accidents involving AVs. This paper first identifies how biases, ethical issues, perverse incentives, and technical issues can affect AVs and then explores how these issues affect algorithmic decision-making in AVs. The paper then examines the potential safety risks, discriminatory outcomes and the factors contributing to liability risks of AV stakeholders stemming from algorithmic decision-making in AVs, discusses the steps taken by several governments to tackle these risks and highlights the outstanding issues that need to be addressed before concluding. The results from our analysis indicate that most of the surveyed governments have begun addressing safety risks. However, while recognising bias and ethical issues arising from the application of AI in AVs, with a few exceptions less has been done to address discriminatory effects and liability concerns.

Keywords: Algorithm, Autonomous vehicle, Driverless car, Ethics, Liability, Biases, Discrimination, Safety, Risk.

[1] Corresponding author: Araz Taeihagh, Lee Kuan Yew School of Public Policy, National University of Singapore, 469B Bukit Timah Road, Li Ka Shing Building, Singapore 259771, spparaz@nus.edu.sg, araz.taeihagh@new.oxon.org

1. Introduction

Algorithms are the building blocks for decision-making in various automated and autonomous systems that increasingly pervade society today in various sectors such as the military, healthcare, and transportation (Arkin et al. 2012; Taeihagh & Lim 2019). Autonomous Vehicles (AVs) rely on algorithms to process data and make decisions more efficiently and accurately than the typical human driver (Riley 2018; Schwarting et al. 2018), potentially reducing the 1.3 million road fatalities that occur worldwide each year that are attributed to human errors (NHTSA 2017; Naughton 2018). The prospect of much higher road safety, efficiency, mobility, and other benefits has prompted many governments to invest heavily in the technology (Pugnetti & Schläpfer 2018; Lim & Taeihagh 2018).

Despite their widespread appeal, algorithmic decision-making in AVs poses several risks to society. Firstly, AVs can introduce new safety risks and potentially perpetuate discrimination in the allocation of risks among different road users. Many studies have warned and proven that data mining processes can result in disparate outcomes and thereby exacerbate discrimination (Barocas & Selbst 2016; Selbst 2017) and in AVs, algorithmic bias can lead the AV to make decisions that prioritise the safety of certain groups of road users over others. Concerns over how AVs should make decisions during unavoidable crashes has prompted proposals of ways to design AVs' algorithms with ethical rules (Liu 2018; Bonnefon et al. 2016), but these approaches can introduce unintended consequences for both safety and discrimination (Goodall 2014; Coca-Vila 2018). Economic incentives of manufacturers and other AV entities also motivate the design of algorithms that shape the AV's decisions to maximise profit rather than safety, and technical issues still limit the AV system's performance. However, limited research has explored AV algorithms' implications for safety

and discrimination. Secondly, it is unclear how responsibilities and liability will be assigned between various AV stakeholders for accidents, exposing AV stakeholders to new liability risks (Martínez-Díaz et al. 2019; Mackie 2018).

Given the above issues and research gaps, this study seeks to analyse how bias, ethics, economic incentives and technical limitations influence algorithmic decision-making processes in ways that can lead to discrimination and increase safety risks to the public. We then highlight how ambiguities in responsibility for algorithmic decisions in AVs and ambiguities in legal liability frameworks lead to liability risks for various AV stakeholders. Lastly, we analyse and compare the steps taken by various governments to address safety risks and liability risks posed by AVs.

2. Background

The Society of Automotive Engineers' (SAE) categorises five levels of vehicle autonomy, starting with levels 1 and 2 where the human driver performs most of the dynamic driving tasks with the assistance of advanced driver assistance systems. At level 3, the vehicle performs all dynamic driving tasks, but the human driver is required to resume control occasionally (SAE 2014; Watzenig & Horn 2017;). At levels 4 and 5, vehicles are considered highly and fully autonomous respectively as they can perform all the driving tasks without human input. Owing to technological advancements in hardware and software systems – Improvements in sensor accuracy enables the vehicle to identify obstacles on the road and respond to unexpected environmental changes (Cho et al. 2014; AV Working Group 2018); Machine-learning (ML) algorithms aided by advancements in computational power, process data about its environment to make driving decisions in real-time; and external

communication networks enable the AV to exchange information with and learn from other connected vehicles and infrastructure (Bathaee 2018; Duarte & Ratti 2018; Lim & Taeihagh 2018). Throughout this study, we focus on vehicles classified under SAE's levels 4 and 5 of autonomy, which we will refer to as "AVs".

The foundation of decision-making in AVs stems from artificial intelligence (AI) through rule-based algorithms, and later included systems with ML capabilities (Bathaee 2018; Osoba & Welser IV 2017). The learning process in ML involves forming mathematical correlations between the input and output data based on pre-programmed decision-making criterion to build an internal model of the world, which is optimised against a new set of data and the model yielding the greatest predictive accuracy is selected (Bathaee 2018). In AVs, ML algorithms are trained using many input variables including weather conditions, traffic signals and the location of obstacles on the road to inform driving decisions to manoeuvre safely on roads (Watzenig & Horn 2017; Coppola & Morisio 2016).

However, the design of AVs' algorithms could yield potentially biased and unethical decisions with consequences for safety and discrimination, which are exacerbated by the presence of economic incentives that motivate various stakeholders in the AV value-chain (Liu 2018; Goodall 2017). Issues in the programming of algorithms are more difficult to recognise and correct in ML than rule-based algorithms as the former model highly complex logics that are not easily understood by humans (Burrell et al. 2016). Safety risks can also arise from software and hardware limitations (Pendleton 2017; Cai et al. 2018). Furthermore, determining liability for accidents remains a challenge due to ambiguities in responsibilities of different AV stakeholders in the value-chain and ambiguities in liability laws when applied to ML systems (Geistfeld 2017; Collingwood 2017). In this study, we examine the

implications of bias, ethics and economic incentives for safety and discrimination, how ambiguities in responsibilities and legal frameworks create liability risks and discuss several steps taken to tackle these issues.

3. AVs, safety risks and discrimination

Through a non-exhaustive but comprehensive examining of the literature on AVs, AI, ML and algorithms in relation to its key societal concerns, we identified bias, ethics, economic incentives, and technical limitations as three key issues that impact algorithmic decision-making in AVs in ways that can lead to biased or unfair allocations of safety risks that result in discrimination, as well as increased safety risks to the public.

3.1. Bias

Bias, a common problem in all computer systems, is referred to as outcomes that "systematically and unfairly discriminate" against certain individuals or groups of individuals in society (Friedman & Nissembaum 1996) and can be introduced in multiple ways in an AV system. Firstly, training data bias may be present when the input data are not statistically representative of the overall population, creating inaccurate classifications and statistically biased outcomes (Danks & London 2017; Lepri et al. 2017). Secondly, bias may result from the selection of sensitive input variables that are not "permitted" by legal and moral standards in certain types of decision-making (Danks & London 2017). When sensitive individual-specific characteristics, such as a person's age, gender and size, are used by the AV as decision-making criteria, the AVs' algorithms may "penalise" and "privilege" some groups of individuals with certain characteristics meet the algorithm's goals, such as minimising the

total quantity of harm,[2] allocating greater safety risks to these individuals during accidents (Liu 2018). Discrimination can be avoided by "arbitrarily and unpredictably" choosing an outcome without considering these characteristics, but doing so may be considered unethical as it involves "choosing between lives without any deliberation" (Lin 2015).

In addition, intentional discrimination may result when algorithms' designers and AV manufacturers intentionally introduce the above forms of bias to benefit from discriminatory outcomes. Decision-makers may exercise their discretion to select an algorithm, distort the training data or select "proxies for protected classes" (Kroll et al. 2016) to satisfy their individual preferences. In particular, manufacturers may be economically incentivised to favour outcomes that are biased towards the safety of AV passengers in order to maximise profits from the sale of AVs and to ensure AVs' commercial success.[3] Safety risks may be disproportionately allocated towards other groups of individuals, and this is exacerbated by the lack of legal frameworks to ensure accountability for the decisions resulting from AVs' algorithmic decision-making processes (Liu 2018).

Several studies have proposed ways to overcome bias and discrimination in autonomous systems in general. Technical tools to remove bias include modifying the data, including anti-discrimination criteria in the algorithm, and modifying algorithmic outputs to reduce or remove the effects of bias on certain groups of individuals (Mittelstadt et al. 2016). The government in South Korea aims to develop such techniques to detect data bias and correct software errors at every stage of AI development (Ministry of Science & ICT 2016) and the UK government will collaborate with the Alan Turing Institute to develop AI talent and

[2] In Section 3.2, we discuss various types of ethical preferences to which AVs may be programmed to follow and their implications of AV safety risks in greater detail.
[3] In Section 3.3, we elaborate on the role of economic incentives in shaping the choice of preferences that are programmed into algorithms in AVs.

auditing tools to mitigate "social inequalities" resulting from algorithmic decision-making (DBEIS 2018).

To mitigate algorithmic bias, principles of explainability and verifiability and the importance of disclosing information on algorithmic decision-making to the public were emphasised in guidelines for AI that were released by governments in Japan and Singapore, where the latter also recommended governance practices in operations management and systems design to increase the accountability of AI-deploying organisations (The Conference towards AI Network Society 2017; PDPC 2018). In the EU, the GDPR prohibits any automated decision that utilises sensitive personal data and that notably affects data subjects and also mandates the right to explanation, by requiring firms to provide data subjects "meaningful", "concise", "intelligible" and "accessible" forms of information regarding the logic behind these decisions (Goodman & Flaxman 2016, p.6). However, scholars highlight that transparency may be insufficient to explain and detect bias in highly complex ML algorithms whose logics are not explicitly programmed and not easily interpretable even by programmers themselves (Guidotti et al. 2018; Lepri et al. 2017).

3.2. Ethics

AVs will have to allocate risks among multiple persons during unavoidable accidents and everyday driving scenarios, such as deciding how much space to give to a nearby cyclist (Pugnetti & Schläpfer 2018; Goodall 2017), both of which involves ethical decision-making. Thus, many scholars have proposed formalising ethical rules[4] and technical approaches to program ethical rules in AVs. To formalise ethical rules, scholars conduct thought experiments such as the trolley problem, where hypothetically, such as in the case that an AV

[4]. The term 'rules' here broadly refers to ethical theories, principles, norms and values.

must decide whether to swerve and crash into one pedestrian or continue in its current path to crash into five pedestrians (Lin 2015). Gathering responses for these situations can reveal insights into the ethical preferences of members of society to inform AVs' decision-making (Goodall 2016; Bonnemains et al. 2018). However, the trolley problem contains several unrealistic assumptions that do not hold in actual driving scenarios, such as assuming complete certainty in the outcomes and that the passenger can choose how harm is distributed as in reality, AVs will face uncertainty and risk (Himmelreich 2018; Goodall 2017).

The second perspective towards the ethics of AVs is concerned with the technical approaches to program ethical rules from specified ethical theories, such as utilitarianism and deontology into the algorithm in a "top-down" approach (Himmelreich 2018; Tzafestas 2018; Wallach et al. 2008). Utilitarianism emphasises on the morality of outcomes and argues that the most ethical action is the one that maximises the total amount of utility or minimises the total quantity of harm (Liu 2018; Johnsen et al. 2017). AV algorithms thus should compute all possible outcomes and actions, the probability and magnitude of consequences and evaluate the "goodness" resulting from each possible outcome (Tzafestas 2018; Taeihagh & Lim 2019). In contrast, deontology emphasises that action, rather than outcomes, should be motivated by respect for all humans, such as Asimov's Three Laws of Robotics (Leenes and Lucivero 2014; Tzafestas 2018), and its rules can be defined in a hierarchical manner, which makes explicit the reasoning behind the algorithm's ethical decision-making (Thornton et al. 2017).

Programming each ethical theory in a top-down fashion introduces their own set of unintended consequences that can undermine AV safety. Firstly, utilitarian-oriented algorithms minimise collective harms and do not consider equity or fairness (Goodall 2014),

which can lead the AV's algorithm to discriminate by allocating more risk to the same groups of individuals and is exacerbated by bias from using sensitive characteristics as decision-making criteria (see Section 3.1). Programming utilitarian principles can also be hindered by existing limitations in machine perception and other technical issues that limit the AV's ability to compute all possible consequences and actions within a short timeframe (Johnsen et al. 2017; Taeihagh & Lim 2019; See Section 3.4). Secondly, algorithms that are programmed to follow hierarchically defined deontological ethical rules may not successfully recommend an appropriate course of action in scenarios where rules conflict or cannot be satisfied, which can hinder the AV's adaptability to new circumstances and create new safety risks, such as instructing the AV to halt in its path (Goodall 2016). Furthermore, deontological rules may not cover all kinds of driving scenarios (Johnsen et al. 2017; Goodall 2014) and are often vague and thus, cannot always be explicitly programmed, such as the definition of "obstruction" or "safe" in different scenarios (Leenes & Lucivero 2014; Goodall 2014). Given each theory's limitations, scholars have advocated combining both utilitarian and deontological ethics to obtain a broad perspective before deciding on the most ethical choice and increasing transparency to justify the system's outcomes (Goodall 2017). Others have proposed a bottom-up approach that allows the AV to learn ethical rules from past data on human ethical judgements during driving, but the algorithm can potentially deviate from its programmed ethical rules as part of the learning process, and the lack of explicitly programmed rules can increase the system's opacity (Danaher et al. 2017; Bonnemains et al. 2018).

Germany released ethical rules for AVs in 2017, whereas several governments have released guidelines, created advisory committees, and are expanding research on AI-related ethical issues. Germany's ethical rules incorporate both utilitarian and deontological principles by

emphasising the prioritisation of human life, yet advocating damage minimisation without discriminating individuals based on personal characteristics (FMTDI 2017). The rules also restrict the use of self-learning systems to non safety-critical functions and highlight the need to publicly disclose the programming of AVs. Japan and Singapore's AI guidelines emphasise on human dignity (The Conference towards AI Network Society 2017), and human centricity (PDPC 2018) and the Japanese government also created an Advisory Board to ensure the sustainability, prosperity, and inclusivity from AI use (Government of Japan 2016). Similarly, South Korea intends to create ethical guidelines for AI (Ministry of Science & ICT 2016).

3.3. Economic incentives

The AV eco-system includes various stakeholders motivated by different incentives, such as AV manufacturers, AV consumers, software designers, ridesharing and transportation network companies, and hardware companies that shape the AV algorithms' decision-making criterion and preferences. AV manufacturers can program the AVs' algorithms to maximise profit rather than to ensure safe driving outcomes, such as to prioritise the safety of AV passengers, which is aligned with customers' incentives to ensure their own safety and thus, causes more safety risks to be allocated to other parties (Liu 2018). Manufacturers can also maintain liability claims at a constant level by programming AV's driving behaviour as a function of average income in a given district, with the implication that AVs would "drive more carefully" in an "affluent" district compared to in an "economically deprived area" (Himmelreich 2018). Doing so could be perceived as discrimination based on income level as it effectively transfers safety risks from areas characterised by higher income levels to areas characterised by lower income levels.

Differing profit incentives of various stakeholders in the AV value-chain also influence the programming of AVs that create systemic effects on the degree of coordination between AVs and other road users, potentially increasing the risk of collisions. Firstly, to differentiate their AVs, manufacturers can configure their algorithms with different decision-making preferences (Himmelreich 2018). The lack of standardisation of the algorithmic preferences can reduce the coordination and predictability of AVs' behaviour and increase risks of collisions. While coordination can be achieved through data sharing between manufacturers, this may be impeded by concerns over privacy and intellectual property rights (Himmelreich 2018). Secondly, stakeholders in the AV value-chain are motivated by different profit measures – AV manufacturers consider the "number of vehicles sold" or the "average margin per vehicle", ridesharing and transportation network companies measure profits based on the "number of trips completed" and average km/mi or margin per km/mi travelled, and data aggregators map mobility data to online behavioural data to maximise the value of the "insights" derived from the data (Neels 2018). The implications of potentially misaligned incentives between AV stakeholders for AV safety have yet to be explored.

3.4. Technical limitations

Various technical limitations still exist in the AV's system that can constrain further improvements in AV technology (Lin et al. 2018). As shown in figure 1, the AV's hardware comprises of sensors for data collection, external communication networks and actuators (e.g. brakes) to execute motor commands, whereas the AV's software components consist of perception, where data is collected from sensors and external communication networks to model its environment; Planning that involves decision-making algorithms optimising over

selected preferences and criterion; and the control component where algorithms compute control inputs for the hardware to execute decisions (González et al. 2016; Pendleton et al. 2017).[5]

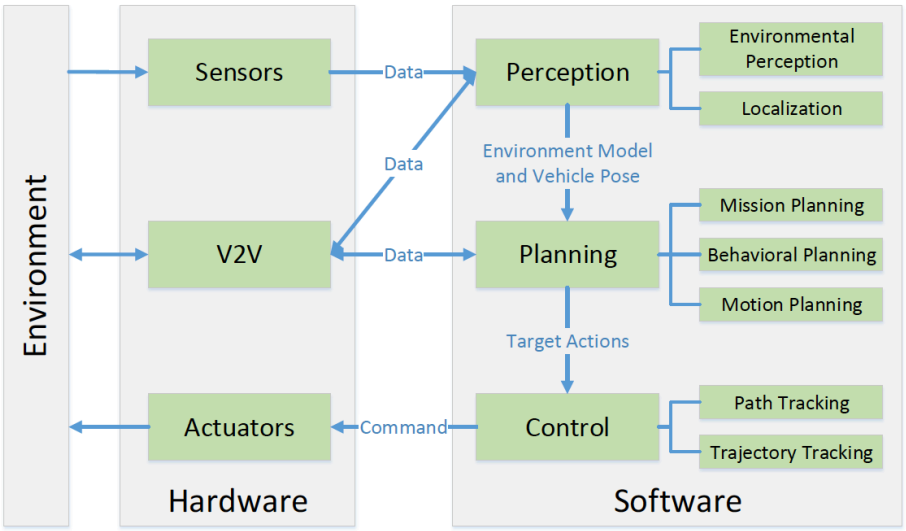

Figure 1: Summary of AV system components (source: Pendleton et al. 2017)

AV perception is constrained by limitations in sensors and ML algorithms' susceptibility to manipulation of image data. Firstly, Global Navigation Satellite System (GNSS)-based sensors are costly, and inaccurate in urban environments due to the obstruction of communication signals by external objects and inconsistencies between HD Map and GNSS coordinates (Cai et al. 2018). While cameras are less costly, they still struggle capturing images in unclear backgrounds, such as during adverse weather (Cai et al. 2018; Koopman & Wagner 2016) and light detection and ranging (LiDAR) sensors are more accurate but costly (Combs et al. 2018). Possible ways to improve sensor accuracy and reduce its costs are combining different sensor types and using ML algorithms to process sensor data to improve object detection (Schwarting et al. 2018; Pendleton et al. 2017). Secondly, sensor input

[5] Other technical issues include Cybersecurity risks and the possibility of hardware failures such as sensor failures due to "electrical failures, physical damage, or age". For a detailed discussion on Cybersecurity concerns in AVs, see Lim & Taeihagh (2018).

images can be artificially modified to force the algorithm to misclassify inputs, such as modifying “STOP” signs to provoke a misclassification and cause the car to continue driving in its path (Papernot et al. 2016; Nguyen et al. 2015). Various methods of defending neural networks against such manipulation have been proposed (e.g. reducing output error fluctuations induced by manipulated inputs (Papernot et al. 2015)), but other proposed ways to detect adversarial samples have been shown to be easily bypassed (Carlini & Wagner 2017).

Furthermore, decision-making algorithms still struggle to understand human interactions and manoeuvre safely around unexpected environmental changes, and control algorithms remain either inaccurate or too computationally expensive. Firstly, it is critical for decision-making algorithms to model interactions with humans in the AV, neighbouring the AV and in other vehicles to drive smoothly while engaging in negotiations such as overtaking and lane changing (Ohn-Bar & Trivedi 2016; Shalev-Shwartz et al. 2016). However, little has been done to explore the implications of these interactions for AV vision and learning capabilities, and scholars highlight the need for more advancements AV sensor resolution, planning methods and in developing standards for modelling these interactions (Schwarting et al. 2018; Ohn-Bar & Trivedi 2016). Secondly, motion planning algorithms may fail to readjust planned trajectories in time when unexpected obstacles emerge (González et al. 2016; Pendleton et al. 2017). To avoid such dangerous scenarios, Pendleton et al. (2017) stress the need for “incremental planning and replanning” to adjust to unexpected environmental changes. In the control component, geometric and kinematic control algorithms only account for the AV’s geometrical dimensions and kinematic properties such as acceleration and velocity, but neglect vehicle dynamics, which can cause vehicle instability and oscillation at high driving speeds (Amer et al. 2017; Dixit et al. 2018). On the other hand, more accurate

control algorithms that account for nonlinear dynamics of the AV remain too computationally costly to be accommodated by the limited computing resources available in existing vehicles (Pendleton et al. 2017; Amer et al. 2017).

4. AVs and liability risks

AV stakeholders in the supply chain shape the programming of its underlying algorithms that replace the human driver in making and executing driving decisions, but how responsibilities for shaping these algorithms and legal liability for harms should be allocated during accidents remains unclear, creating liability risks for AV stakeholders, which will be examined in this section.

4.1. Ambiguities in responsibilities

Amidst the debate in the legal and philosophical literature, there is a lack of consensus regarding whether AV manufacturers, algorithm programmers or the human AV passenger should be responsible for AV accidents and what their responsibilities are. Firstly, unlike human drivers who are forced to react instinctively to unexpected accident scenarios, AV manufacturers and programmers design these algorithmic decision-making processes without any time-constraints and thus, should arguably be expected to make "better decisions" than human drivers (Pugnetti & Schläpfer 2018; Lin 2015). However, programmers' and manufacturers' capacity to control the AV's resulting behaviour is limited due to the unpredictability of the autonomous system's decision-making, particularly of ML algorithms (Liu 2017; Nyholm 2018). Furthermore, holding manufacturers responsible for AV accidents can introduce excessive liability risks for manufacturers, which many have recognised can

dampen longer-run innovation and safety improvements (Leenes and Lucivero 2014; Nyholm 2018).

Studies argue that instead of holding manufacturers responsible, allocating some responsibility to the AV user to stay alert and intervene in anticipated accidents would be beneficial for longer-run technological improvements (Hevelke & Nida-Rümelin 2015). However, accidents or harms can still result despite AV users fulfilling certain pre-specified obligations and AV manufacturers and programmers, ensuring the absence of any manufacturing and design defects (Liu 2017). Humans also take a long time to regain situational awareness to effectively intervene, which can be further delayed by factors such as age, physical or psychological impairments, boredom, or fatigue (Hevelke & Nida-Rümelin 2015). Alternatively, scholars recommend focusing more on system design, from which regulators should learn from and enforce greater safety improvements (Winfield & Jirotka 2017) or holding all AV users "collectively responsible" for AVs' safety risks through a mandatory tax or insurance (Nyholm 2018; Hevelke & Nida-Rümelin 2015).

Ambiguities in the responsibilities of all stakeholders in the AV value-chain can potentially lead AV manufacturers to escape or to displace liability onto AV users. Firstly, under existing legal liability frameworks, the AV itself may be held "causally responsible" for the outcome whereby the harms from the AV accident are simply categorised as "damages" and treated as a "natural phenomena" without ethical or legal consequences, with the possibility of using moral and legal interpretation to defer or prevent the ascribing of responsibility (Liu 2017). Secondly, ambiguities around responsibility can result in "scapegoating" whereby manufacturers seek to reduce potential liability claims by displacing liability onto AV users (Liu 2018). As AV users are not able to shape the AV's driving decisions but simultaneously

can be held responsible for their failure to intervene, manufacturers can easily push responsibility to humans, such as equipping AVs with "warnings" or ways of "reverting control" (Liu 2018; Elish 2016)

Possible ways to clarify responsibility allocation to mitigate the above unintended consequences include clarifying the responsibilities of AV stakeholders based on their roles, the rights they enjoy over AVs and the ways in which they exercise control over AVs (Nyholm 2018). On the other hand, Liu (2017) argues that instead of focusing on ideas of capabilities, capacity, "control and causation", responsibility doctrines should focus on holding the beneficiaries of AVs, which comprise of the AV users, manufacturers and programmers, accountable for the risks they impose on third parties in society.

Several countries have issued public consultations, policy recommendations and guidelines for industry to clarify various AV stakeholders' new responsibilities. After gathering feedback from consultations, Australia's National Transport Commission (NTC) issued policy recommendations to clarify the distribution of legal responsibilities (NTC 2018), which includes creating or amending new driving laws that ensure the existence of and clarifies who is the responsible legal entity, and identifies obligations of the AV system and of a "fallback-ready user" to stay alert to respond to emergencies and regain control of the AV if necessary (NTC 2018).[6] Germany's ethical rules for AVs (see Section 3.2) clarifies the responsibilities of a range of parties that operate components of the AV and connected traffic infrastructure. For instance, entities that operate the original equipment manufacturer's

[6] The national working group intends to further examine and reach a consensus on these new obligations, after which the NTC will consult with key industry stakeholders and start drafting legislation after May 2019.

backend IT systems are responsible for the quality and reliability of AV data, whereas telecommunications operators are responsible for secure data transmission (FMTDI 2017).

4.2. Legal ambiguities

Existing tort liability regimes[7] can subject AV manufacturers, algorithm programmers and other entities in the AV supply chain to products liability based on strict products liability for product defects (Glancy 2015), which includes manufacturing defects and/or design defects, but there are several issues in determining the existence of defects and the degree to which the AV manufacturer or algorithm programmer caused these defects. Firstly, an AV is considered to have manufacturing defects when it did not operate as intended by the manufacturer, such as when a sensor's poor wiring resulted in malfunctions and caused an accident (Kim et al. 2017; Mackie 2018). Determining the AV manufacturer's liability requires both proof that a malfunction existed and that the manufacturer caused the defect, which is challenging as the defect could have been introduced by other third parties after the manufacturing process, such as during repair works (Mackie 2018), and different interpretations of product malfunctions could be used to argue that the AV system "proximately" caused the accident (Geistfeld 2017).

Secondly, strict product liability can hold AV manufacturers and programmers liable for design defects in the AV software, such as coding errors, that "proximately caused" the AV accident (Geistfeld 2017). Courts typically use two types of tests to determine design defects that contain several limitations when applied to ML systems – the "risk utility test" assumes

[7] AVs introduce concerns around both civil and criminal liability. For the brevity of this study, we focus on civil liability, specifically based on tort liability that comprises of strict product liability based on product defects or on negligence (Glancy 2015).

that only rule-based algorithms shape the system's behaviour, but ML systems depend largely on learnt rules from the data; and the "consumer expectations test" defines a design defect as one that yields "unreasonably dangerous" outcomes beyond an "ordinary" consumer's expectations,[8] which may be inappropriate as consumers' lack sufficient technical knowledge on AVs and their expectations can change significantly over time (Geistfeld 2017; Kim et al. 2017). The extent to which the programmer vis-à-vis the manufacturer should have anticipated certain software errors and the extent of testing that manufacturers should have conducted to reveal possible erroneous behaviours in advance remains unclear (Kim et al. 2017; Mackie 2018).

Product liability based on negligence could be imposed on entities for their failure to devote sufficient attention in the manufacture, distribution or sale of a product to mitigate potentially avoidable harm caused to others (Gless et al. 2016; Glancy 2015). Even without a manufacturing or design defect, AV manufacturers and programmers can be held liable for failing to design AVs to "protect" or at least be "impartial" to the interests of bystanders, such as programming the AV to treat both consumers and bystanders equally or to be capable of communicating with pedestrians (Geistfeld 2017), as well as equipping the AV with sufficient warnings and instructions to inform consumers and bystanders of the safety and risks involved in using the AV (Kim et al. 2017; Mackie 2018). Thus, scholars have recommended that manufacturers can avoid these liability risks by equipping the AV with sufficient warnings before the sale of the AV, particularly if a "foreseeable risk" of harms that would significantly influence the consumer's decision to use the AV was found, as well as after the sale through the AV's online notifications (Geistfeld 2017).

[8] To be "unreasonably dangerous", the product "must be dangerous to an extent beyond that which would be contemplated by the ordinary consumer who purchases it, with the ordinary knowledge common to the community as to its characteristics" (Kim et al. 2017).

To address liability risks from AVs, a "first party" automobile insurance (or "no fault" liability regime) has been proposed and implemented by several governments, which removes the need for accident victims to engage in costly lawsuits to prove the fault of another party and ensures compensation for victims by the "first party insurer" who will take recourse against another party (Eastman 2016; Vellinga 2017). In 2017, the UK government passed the Bill HC 143 that requires insurance companies to automatically compensate accident victims for most of the damages caused by insured AVs under the existing motor vehicle insurance scheme (Taeihagh & Lim 2019). Similarly, the EU evaluated their Motor Insurance Directive as sufficient to ensure compensation to AV accident victims, whereby insurers can take recourse against manufacturers under the existing Product Liability Directive (EC 2018). Other recommendations to manage liability risks include increasing transparency, such as installing "black boxes" or event data recorders into AVs to aid investigations into the decision-making of AVs before and during an accident, which was mandated by Germany's new AV legislation in June 2017 and recommended by the Japanese government in 2018 (Taeihagh & Lim 2019). Scholars also recommend standardising and certifying black box designs (Winfield and Jirotka 2017) and increasing the interpretability of algorithms, but the latter remains challenging for highly opaque algorithms (McAllister et al. 2017; see Section 3.1).

5. Discussion and conclusions

This study examined several key factors in algorithmic decision-making in AVs that influence the AV's driving decisions, their implications for safety and discrimination and several steps taken to address them. Bias may be unintentionally introduced through the AVs'

data, choice of decision-making criterion and algorithms and intentionally introduced by AV stakeholders to prioritise the safety of certain groups of individuals to maximise profits. To identify discrimination and bias, transparency can be increased by ensuring the traceability of outputs to inputs and by increasing algorithmic interpretability, but this remains a challenge due to the opacity of complex ML algorithms. Singapore and Japan have issued AI guidelines not specific to AVs that emphasise on explainability and transparency of algorithms' programming, while the EU GDPR prohibits algorithmic decision-making that utilises sensitive personal data and mandates a "right to explanation". Further steps have yet to be taken to analyse biases and their implications for safety and discrimination in the context of AVs.

AVs can also introduce new safety risks and discrimination from incorporating ethical rules in AVs' algorithms. Trolley problems contain unrealistic assumptions that are incompatible with actual driving scenarios; the top-down approach of programming utilitarian ethical principles that aim to minimise harms can consistently allocate more risks to particular individuals, yielding discrimination, and are limited by computational limitations in the AV system, whereas deontological ethical rules can conflict, may not be sufficiently comprehensive and may not be possible to explicitly program; Learnt ethical rules from the bottom-up approach can be potentially unethical and increase the system's opacity. Several governments have created AI ethical guidelines and committees, whereas Germany has released AV-specific ethical guidelines that combine utilitarian and deontological principles, but more research to examine and reconcile trade-offs in programming different ethical principles into AVs is required.

Economic incentives motivate key AV stakeholders to program AV's driving decisions in ways that potentially undermine road safety and yield discrimination, such as manufacturers' prioritisation of AV passengers' safety to maximise profit which can yield unfair allocations of safety risks to others, and differentiating AVs' algorithmic preferences that yield reduced traffic coordination and thus, greater safety risks. The implications of the interactions and potential conflicts in incentives among other third parties in the AV value-chain also requires more examination in future work.

Various technical issues can undermine the AV system's performance and yield dangerous driving behaviour. In AV perception, sensors remain either too costly or inaccurate during harsh driving conditions, and adversaries can easily manipulate input images to cause misclassifications, which remain difficult to detect. In addition, more research has yet to examine issues in modelling human-vehicle interactions in AVs' decision-making algorithms, adjusting planned trajectories to unexpected environmental changes and their implications for road safety. Lastly, different control algorithms face trade-offs in accuracy and computational cost, which necessitates more research to improve algorithms' efficiency and greater advancements in processor speeds and memory (Pendleton et al. 2017).

Lastly, this study examined the factors that contribute to liability risks for AV stakeholders. Firstly, ambiguities in the responsibilities of AV stakeholders can undermine the assignment of liability. Holding AV manufacturers and programmers responsible can introduce significant liability risks that can dampen further safety improvements in AVs and motivate manufacturers to displace liability onto AV passengers for failing to sufficiently monitor the AV and intervene in accidents, but the latter may be unethical as humans' interventions may be ineffective or inappropriate. Ways to clarify responsibilities have been proposed, such as

analysing the roles and benefits enjoyed by AV stakeholders. Australia's government has consulted the public and recommended steps to identify responsibilities for the AV system and human driver to be incorporated in future laws, and Germany's ethical rules for AVs outlines the responsibilities of other stakeholders that operate connected traffic infrastructure. Secondly, under strict product liability regimes, it is challenging to determine manufacturing and software design defects and the cause of these defects, as AV stakeholders have limited control over the unpredictable and data-driven decisions of ML systems, and AV stakeholders can also be held liable for negligence. These risks can be avoided by providing AV users with sufficient warnings, by adopting no fault liability regimes that ensure compensation to AV victims without requiring proof of fault, which has been implemented in the UK and EU, and/or by mandating black boxes in AVs, which has been implemented in Germany and recommended in Japan.

References

Amer, N. H., Zamzuri, H., Hudha, K., & Kadir, Z. A. (2017). Modelling and control strategies in path tracking control for autonomous ground vehicles: a review of state of the art and challenges. *Journal of Intelligent & Robotic Systems, 86*(2), 225-254.

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE*, 100(3), 571-589.

Autonomous Vehicles (AV) Working Group. (2018). Report of the Massachusetts Autonomous Vehicles Working Group. https://www.mass.gov/files/documents/2018/09/12/DraftReport_AV_WorkingGroup.pdf

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Cal. L. Rev.*, *104*, 671.

Bathaee, Y. (2018). The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law & Technology*, *31*(2), 889.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science, 352(6293), 1573-1576.

Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. Ethics and Information Technology, 20(1), 41-58.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 2053951715622512.

Cai, H., Hu, Z., Huang, G., Zhu, D., & Su, X. (2018). Integration of GPS, Monocular Vision, and High Definition (HD) Map for Accurate Vehicle Localization. *Sensors, 18*(10), 3270.

Carlini, N., & Wagner, D. (2017, November). Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* (pp. 3-14). ACM.

Cho, H., Seo, Y. W., Kumar, B. V., & Rajkumar, R. R. (2014, May). A multi-sensor fusion system for moving object detection and tracking in urban driving environments. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on* (pp. 1836-1843). IEEE.

Coca-Vila, I. (2018). Self-driving cars in dilemmatic situations: An approach based on the theory of justification in criminal law. Criminal Law and Philosophy, 12(1), 59–82.

Collingwood, L. (2017). Privacy implications and liability issues of autonomous vehicles. Information & Communications Technology Law. 26(1):32–45.

Combs, T. S., Sandt, L. S., Clamann, M. P., & McDonald, N. C. (2018). Automated vehicles and pedestrian safety: exploring the promise and limits of pedestrian detection. *American journal of preventive medicine*.

Coppola, R., & Morisio, M. (2016). Connected car: Technologies, issues, future trends. *ACM Computing Surveys (CSUR)*, 49(3), 1-36. doi:10.1145/2971482

Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., ... & Murphy, M. H. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. Big Data & Society, 4(2), 2053951717726554.

Danks, D., & London, A. J. (2017, August). Algorithmic bias in autonomous systems. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (pp. 4691-4697).

DBEIS, Centre for Connected and Autonomous Vehicles, and Jesse Norman MP. (2018). Government to review driving laws in preparation for self-driving vehicles. *Crown*. https://www.gov.uk/government/news/government-to-review-driving-laws-in-preparation-for-self-driving-vehicles

Duarte, F., & Ratti, C. (2018). The Impact of Autonomous Vehicles on Cities: A Review. Journal of Urban Technology, 1-16.

Eastman, A. D. (2016). Is No-Fault Auto Insurance the Answer to Liability Concerns of Autonomous Vehicles? American Journal of Business and Management, 5(3), 85-90.

Elish, M. C. (2016). Moral crumple zones: Cautionary tales in human–robot interaction. WeRobot 2016.

European Commission (EC). (2018). Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee, The Committee of the Regions. On the road to automated mobility: An EU strategy for mobility of the future. Brussels, 17.5.2018

Federal Ministry of Transport and Digital Infrastructure (FMTDI). (2017). Ethics Commission Automated and Connected Driving. Federal Ministry of Transport and Digital Infrastructure Report.

https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems (TOIS), 14(3), 330-347.

Geistfeld, M. A. (2017). A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation. Cal. L. Rev., 105, 1611.

Glancy, D.J. (2012). Privacy in Autonomous Vehicles. *Santa Clara Law Review,* 52(4):1171–1239.

Gless, S., Silverman, E., & Weigend, T. (2016). If Robots cause harm, who is to blame? Self-driving Cars and Criminal Liability. New Criminal Law Review: In International and Interdisciplinary Journal, 19(3), 412-436.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine, 38(3), 50-57.

Government of Japan. (2016). The 5th Science and Technology Basic Plan. Cabinet Office, Government of Japan. https://www8.cao.go.jp/cstp/english/basic/5thbasicplan.pdf

González, D., Pérez, J., Milanés, V., & Nashashibi, F. (2016). A Review of Motion Planning Techniques for Automated Vehicles. *IEEE Trans. Intelligent Transportation Systems, 17*(4), 1135-1145.

Goodall, N. J. (2014). Machine ethics and automated vehicles. In Road vehicle automation (pp. 93-102). Springer, Cham.

Goodall, N. J. (2016). Away from trolley problems and toward risk management. Applied Artificial Intelligence, 30(8), 810-821.

Goodall, N. J. (2017). From trolleys to risk: Models for ethical autonomous driving. American Journal of Public Health, 107(4), 496-496. doi:10.2105/AJPH.2017.303672

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*(5), 93.

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: An ethical analysis. Science and Engineering Ethics, 21(3), 619-630. doi:10.1007/s11948-014-9565-5

Himmelreich, J. (2018). Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations. Ethical Theory and Moral Practice, 1-16.

IEEE Standards Association. (2018). Ethically Aligned Design, Version 2. IEEE. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

Johnsen, A., Kraetsch, C., Možina, K., & Rey, A. D2. 1 Literature review on the acceptance and road safety, ethical, legal, social, and economic implications of automated vehicles.

Kim, S. (2017). Crashed Software: Assessing Product Liability for Software Defects in Automated Vehicles. *Duke L. & Tech. Rev.*, *16*, 300.

Koopman, P., & Wagner, M. (2016). Challenges in autonomous vehicle testing and validation. *SAE International Journal of Transportation Safety, 4*(1), 15-24.

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. U. Pa. L. Rev., 165, 633.

Leenes, R., & Lucivero, F. (2014). Laws on robots, laws by robots, laws in robots: regulating robot behaviour by design. Law, Innovation and Technology, 6(2), 193-220.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2017). Fair, Transparent, and Accountable Algorithmic Decision-making Processes. Philosophy & Technology, 1-17.

Lin, P. (2015). Why ethics matters for autonomous cars. In Autonomes fahren (pp. 69-85). Springer Vieweg, Berlin, Heidelberg.

Lin, S. C., Zhang, Y., Hsu, C. H., Skach, M., Haque, M. E., Tang, L., & Mars, J. (2018, March). The architectural implications of autonomous driving: Constraints and acceleration. In *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems* (pp. 751-766). ACM.

Lim, H.S.M.; Taeihagh, A. Autonomous Vehicles for Smart and Sustainable Cities: An In-Depth Exploration of Privacy and Cybersecurity Implications. Energies 2018, 11, 1062. https://doi.org/10.3390/en11051062

Liu, H. (2017). Irresponsibilities, inequalities and injustice for autonomous vehicles. Ethics and Information Technology, 19(3), 193-207.

Liu, H. Y. (2018). Three Types of Structural Discrimination Introduced by Autonomous Vehicles. University of California Davis Law Review Online, 51, 149 – 180. https://lawreview.law.ucdavis.edu/online/vol51/51-online-Liu.pdf

Martínez-Díaz, M., Soriguera, F., & Pérez Pérez, I. (2019). Autonomous driving: a bird's eye view. *IET Intelligent Transport Systems*, *13*(4), 563-579.

McAllister, R., Gal, Y., Kendall, A., Van Der Wilk, M., Shah, A., Cipolla, R., & Weller, A. (2017, August). Concrete problems for autonomous vehicle safety: advantages of Bayesian deep learning. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (pp. 4745-4753). AAAI Press.

Ministry of Science & ICT. (2018). The Innovation Growth Engine Leading preparations for the Fourth Industrial Revolution. Ministry of Science & ICT, Korea. https://english.msit.go.kr/cms/english/pl/policies2/__icsFiles/afieldfile/2018/04/06/%ED%98%81%EC%8B%A0%EC%84%B1%EC%9E%A5%EC%98%81%EB%AC%B8-%EC%9D%B8%EC%87%84%EB%B3%B8.pdf

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).

Naughton, K. (2018). Just How Safe Is Driverless Car Technology, Really? Bloomberg. https://www.bloomberg.com/news/articles/2018-03-27/just-how-safe-is-driverless-car-technology-really-quicktake

NHTSA. (2017). 2016 fatal motor vehicle crashes: overview. Traffic safety facts research note, 2017, 1-9.

NTC. (2018). Changing driving laws to support automated vehicles. National Transport Commission. http://www.ntc.gov.au/Media/Reports/(B77C6E3A-D085-F8B1-520D-E4F3DCDFFF6F).pdf

Neels, C. (2018). The Importance of Incentives in the Development of Autonomous Vehicles. Medium. https://medium.com/@cneels/the-importance-of-incentives-in-the-development-of-autonomous-vehicles-967409458597

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 427-436).

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on Human–Robot collaborations and responsibility-loci. Science and Engineering Ethics, 24(4), 1201-1219. doi:10.1007/s11948-017-9943-x

Ohn-Bar, E., & Trivedi, M. M. (2016). Looking at humans in the age of self-driving and highly automated vehicles. *IEEE Transactions on Intelligent Vehicles, 1*(1), 90-104.

Osoba, O. A., & Welser IV, W. (2017). An intelligence in our image: The risks of bias and errors in artificial intelligence. Rand Corporation.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016, March). The limitations of deep learning in adversarial settings. In Security and Privacy (EuroS&P), 2016 IEEE European Symposium on (pp. 372-387). IEEE.

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2015). Distillation as a defence to adversarial perturbations against deep neural networks. *arXiv preprint arXiv:1511.04508*.

Pendleton, S. D., Andersen, H., Du, X., Shen, X., Meghjani, M., Eng, Y. H., ... & Ang, M. H. (2017). Perception, planning, control, and coordination for autonomous vehicles. *Machines, 5(*1), 6.

Personal Data Protection Commission Singapore (PDPC). 5 June, 2018. Discussion paper on artificial intelligence (AI) and personal data – Fostering responsible development and adoption of AI. *PDPC*. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/Primer-for-Discussion-Paper-on-AI-and-PD---050618.PDF?la=en

Pugnetti, C., & Schläpfer, R. (2018). Customer Preferences and Implicit Trade-offs in Accident Scenarios for Self-Driving Vehicle Algorithms. *Journal of Risk and Financial Management*, 11(2), 28.

Riley, A. The algorithm at the heart of autonomous truck safety. Medium. https://medium.com/@antonyriley/the-algorithm-at-the-heart-of-autonomous-truck-safety-5be09203e5dc

Santos-Lang, C. (2002). Ethics for Artificial Intelligences. In Wisconsin State-Wide technology Symposium "Promise or Peril?". Wisconsin, USA: Reflecting on computer technology: Educational, psychological, and ethical implications.

Schellekens, M. (2015). Self-driving cars and the chilling effect of liability law. Computer Law & Security Review, 31(4), 506-517.

Schwarting, W., Alonso-Mora, J., & Rus, D. (2018). Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*.

Selbst, A. D. (2017). Disparate Impact in Big Data Policing. *Ga. L. Rev.*, *52*, 109.

Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*.

SAE International. (2014). International Standard J3016 Taxonomy and Definitions for Terms related to On-Road Motor Vehicle Automated Driving Systems. https://www.sae.org/misc/pdfs/automated_driving.pdf

Taeihagh, A., & Lim, H. S. M. (2019). Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. Transport reviews, 39(1), 103-128. https://doi.org/10.1080/01441647.2018.1494640

The Conference toward AI Network Society. (2017). Draft AI R&D GUIDELINES for International Discussions. The Conference toward AI Network Society. Japan. http://www.soumu.go.jp/main_content/000507517.pdf

Thornton, S. M., Pan, S., Erlien, S. M., & Gerdes, J. C. (2017). Incorporating ethical considerations into automated vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, *18*(6), 1429-1439.

Tzafestas, S. (2018). Roboethics: Fundamental Concepts and Future Prospects. Information, 9(6), 148.

Vellinga, N. E. (2017). From the testing to the deployment of self-driving cars: legal challenges to policymakers on the road ahead. Computer Law & Security Review, 33(6), 847-863.

Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. Ai & Society, 22(4), 565-582. doi:10.1007/s00146-007-0099-0

Watzenig, D., & Horn, M. (2017). Introduction to automated driving. In *Automated Driving* (pp. 3-16). Springer, Cham.

Winfield, A. F., & Jirotka, M. (2017, July). The case for an ethical black box. In Conference Towards Autonomous Robotic Systems (pp. 262-273). Springer, Cham.