Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Autonomous Vehicles and Embedded Artificial Intelligence: The Challenges of Framing Machine Driving Decisions

Martin Cunneen, Martin Mullins & Finbarr Murphy

To cite this article: Martin Cunneen, Martin Mullins & Finbarr Murphy (2019) Autonomous Vehicles and Embedded Artificial Intelligence: The Challenges of Framing Machine Driving Decisions, Applied Artificial Intelligence, 33:8, 706-731, DOI: 10.1080/08839514.2019.1600301

To link to this article: https://doi.org/10.1080/08839514.2019.1600301

Published online: 13 May 2019.

APPLIED ARTIFICIAL INTELLIGENCE
2019, VOL. 33, NO. 8, 706–731
https://doi.org/10.1080/08839514.2019.1600301

Taylor & Francis
Taylor & Francis Group

Autonomous Vehicles and Embedded Artificial Intelligence: The Challenges of Framing Machine Driving Decisions

Martin Cunneen, Martin Mullins, and Finbarr Murphy

University of Limerick

ABSTRACT

With the advent of autonomous vehicles society will need to confront a new set of risks which, for the first time, includes the ability of socially embedded forms of artificial intelligence to make complex risk mitigation decisions: decisions that will ultimately engender tangible life and death consequences. Since AI decisionality is inherently different to human decision-making processes, questions are therefore raised regarding how AI weighs decisions, how we are to mediate these decisions, and what such decisions mean in relation to others. Therefore, society, policy, and end-users, need to fully understand such differences. While AI decisions can be contextualised to specific meanings, significant challenges remain in terms of the technology of AI decisionality, the conceptualisation of AI decisions, and the extent to which various actors understand them. This is particularly acute in terms of analysing the benefits and risks of AI decisions. Due to the potential safety benefits, autonomous vehicles are often presented as significant risk mitigation technologies. There is also a need to understand the potential new risks which autonomous vehicle driving decisions may present. Such new risks are framed as decisional limitations in that artificial driving intelligence will lack certain decisional capacities. This is most evident in the inability to annotate and categorise the driving environment in terms of human values and moral understanding. In both cases there is a need to scrutinise how autonomous vehicle decisional capacity is conceptually framed and how this, in turn, impacts a wider grasp of the technology in terms of risks and benefits. This paper interrogates the significant shortcomings in the current framing of the debate, both in terms of safety discussions and in consideration of AI as a *moral* actor, and offers a number of ways forward.

> The self-driving car raises more possibilities and more questions than perhaps any other transportation innovation…self-driving cars have become the archetype of our future transportation. Still, important concerns emerge. Will they fully replace the human driver? What ethical judgments will they be called upon to make? What socioeconomic impacts flow from such a dramatic change? Will they disrupt the nature of privacy and security?

CONTACT Martin Cunneen martin.cunneen@ul.ie

(NHTSA, 2016)

Part One: Introduction

Artificial Driving Intelligence: Context to the Significance of Autonomous Vehicle Decisions

Autonomous vehicles (AV) offer the opportunity to harness the benefits of the latest sensory technologies and artificial intelligence (AI) to make driving decisions which mitigate many risks associated with human driving decisions. Indeed, the focus on the AI driving of AV gives rise to two contrasting formulations of the decisional benefits and risks of the technology which epitomise general disorientation regarding the issues of machine decisionality and impacts in terms of benefits and risks to society as a whole. The combination of sensory and intelligence technologies provides a topographical representation of the road phenomenon which support AI to make more immediate and accurate driving decisions. As machines, AV also eliminate decisionality problems associated with the human frailties of fatigue, misperception, and intoxication, along with the problematic decisions humans often make in the context of driving. This rendering of the technological benefits of AV constitutes a safety argument that not only identifies the welfare benefits of machine decisions but also endorses claims that AV should be supported by policy. Conversely, the alternative perspective highlights these potential new risks as decision errors and limitations to the driving AI. As such, there is a clear need to define and disseminate the benefits of AV decisional intelligence in order to avoid underutilisation of the technology due misplaced risk perception (Floridi, 2018).

Governance and Framing of AI Decisions

The roll-out of an emerging technology presents numerous challenges for governance regimes, while the transitioning of such technologies from nanomaterial to GMO entails the assessment and understanding of the benefits and risks. It is therefore crucial to anticipate possible risks. AV are an topical example of a socially embedded and potentially ubiquitous AI technology. Here, as with other technologies, those charged with risk governance regimes face the dichotomy of empirical objectivity and moral dilemma. While there are clearly persuasive utilitarian AV safety arguments predicated on the reduction of deaths and injuries, other moral questions emerge which are more problematic and emphasise the need to exercise caution in introducing AV. Such debates address the fundamental premise of the desirability of machine decisionality over matters of human life and death decisions and mainly obtain to instances whereby automated vehicles confront *moral* choices in the course

of a journey. At one end of the spectrum, AV choices could optimise route-planning algorithms to help avoid schools or known problematic areas, while at the other end, such choices could influence possible road traffic accident (RTA) decisions. Such split-second pre-collision decision windows are thus contextualised in scenarios of unavoidable traffic accidents which may result in death and injury. This paper therefore reflects on the balance between risk and utility inherent in moral questions. As such, it considers both sets of arguments and advances the case for more focus and precision in the conceptual framing of such debates. While this paper acknowledges that AVs are likely to provide a safer form of driving than human drivers in the long-term, it nonetheless interrogates the shortcomings of both empirically-based safety arguments and the various ethical nuances. To better understand the full compass of AI decisionality, it is necessary to weigh claims that AV can mitigate human driving risks by making safer driving decisions against claims that all AI decisionality is circumscribed by an inherent absence of moral agency. In both instances, a clearer elucidation of the decisional capabilities of AI in the context of AV would offer greater clarity.

Autonomous Vehicle Literature Space

Autonomous vehicles offer many significant societal benefits: from enhancing the mobility of those who lack it, transforming urban spaces and supporting the environment, to radically improving safety and saving lives. However, since the opportunities of any substantive technology also carry both embedded and new forms of risk, any actualisation of potential AV benefits also necessitates mitigation of the risks involved. Moreover, AV risk mitigation cannot be undertaken by state governance regimes alone, but must rather be a multi-stakeholder phenomenon. In this instance, traditional state government and new governance models are simply outpaced, as is evident throughout the current era of digital innovation (Marchant, 2011), and highlighted by the NHTSA (2017); "*the speed with which HAVs are advancing, combined with the complexity and novelty of these innovations, threatens to outpace the Agency's conventional regulatory processes and capabilities*." For these reasons, intelligence technologies can only be responded to by a shared risk mitigation process wherein numerous actors cooperate. As such, the conceptualisation and framing of technology in terms of meaning, benefits, and risks, will ultimately determine how stakeholders engage with the technology.

The key consideration of AV risk mitigation discussed across the literature concerns assessment of the AV capacity to make driving decisions. As such, any research which further illuminates the decisionality phenomenon of AV both contributes to the multi-stakeholder risk mitigation process and promotes access to AV societal benefits. Moreover, analysis of the scope of AV decisions in terms of both benefits (risk mitigation) and potential limitations (new forms

of risk) supports the dynamics of new governance relations which consist of both top-down and bottom-up movement. It is of further interest to consider that while AVs arguably afford opportunities to minimise and potentially eliminate the many risks associated with human driving, future benefits cannot be realised unless accurate and effective anticipatory risk governance research is undertaken today. A broad and immensely complex decisional context is inherent to AV technologies, such as, investigating how different governance actors and policy-writers understand the decisional capacity and societal impact of AV decisions (Allenby, 2011; Anderson et al., 2014; Lin, 2013). It also concerns diverse ethical interpretations of AV decisions, including those who identify the need to control ethical decisions as a predetermined configuration of action (Gogoll and Müller, 2016) or calculating metrics such as values and risk (Bonnefon et al., 2015; Goodall 2016; Young 2016). In addition, considerable extant research rehearses the many abstract questions surrounding machine ethics (Malle, 2013; Lin, 2016; Dogan *et al.*, 2015; Gerdes & Thornton, 2016), while others consider meaning, conceptual confusions, and limited decisional capacity (SAE, 2016; Nyholm & Smids, 2016). Issues of the technical scalability of AV decisional capacity are also of significance (Fraichard, 2014), along with a layer of legality and governance which is heavily reliant on fulsome understanding and anticipation of the impacts of AV, particularly in terms of societal risk (Loh and Loh, 2017; Kasperson *et al.*, 2009; Hevelke & Rümelin, 2015).

This journal has previously highlighted such issues. For instance, Trappl (2016) underscores the need to consider the important conceptual differences between human and machine contexts of moral decisionality in the context of AV, while Bringsjord and Sen (2016) highlight the potential confusion surrounding the differing contexts of intelligence and ethical capacities of AV. They also point out the vital need to support actors in reaching more accurate and informed choices in terms of AV policy and regulation. Millar (2016) proposes the need to investigate ethical decision-making in AV, while Goodall (2016) shifts the emphasis from ethics to risk. Others, such as Coeklebergh (2016), attempt to elucidate the importance of changes in relations which socially embedded technologies bring about between agents and actors (Coeklebergh, 2016). This is most evident in the consideration of key legal and ethical concepts by way of the changing human phenomenological relations regarding AV (Coeckelbergh, 2016). However distinct these approaches, they are united in their attempts to fathom the decisional relations of AI and applied applications such as AV. This paper therefore seeks to interrogate how AI and AV as decision-making technologies are conceptually framed and how such framings determine the engagement and understanding of diverse agents (Burgess and Chilvers, 2006; Coates and Coates, 2016). It is our contention that this causal, societal chain of understanding AV

decisions rests upon accurate conceptual (re)framing of the technology as it continues to emerge and evolve. Understanding the role of AV decisionality is itself a complex challenge which requires careful elucidation (Choudury, 2007; Young, 2016), while the basic function of AV requires the driving intelligence to make decisions affecting human welfare and life (Lin, 2016). In fact, AV will typically make thousands of such decisions on every trip, and global deployment will translate into millions of such decisions per day. Accordingly, it is imperative to explore the many facets of the AV decisional spectrum, not merely on terms of awareness of the limitations of AV decisionality, but also with a knowledge of the key contexts wherein different actors confuse or misunderstand the meaning of AV decisions.

Aims

The paper holds the AV decisional phenomenon to consist of two conceptual frameworks regarding decision capacity and risk mitigation. The term *conceptual framework* refers to the way AV and AI technologies are currently presented in terms of decisional capacity and risk mitigation. Thus, the decisional phenomenon of the technologies is delineated in two forms which contrast to human driving decisions. In the first instance, AI in the context of driving is held to offer superior decisions which mitigate risks; in the second, the phenomenon of AI as a driving intelligence presents decisional limitations which present new risks which reside in AI capacity to make moral determinations. While many pressing questions arise in the literature regarding how to best investigate and anticipate the societal impact of emerging technologies in terms of risk (Asveld and Roeser, 2009), conspicuously little attention is given to conceptual analyses, particularly in terms of AV conceptualisation frameworks. Asveld and Roeser (2009), for example, emphasise not only a deficiency in investigating technological risk and morality, but also an apparent gap in applying analytical methods to the emerging technological risk phenomenon. This is most apparent when one considers the need for enhanced transparency and explainability necessary to enable the diverse array of stakeholders to contribute to issues pertaining to the societal use of AI technologies. This is also crucial in terms of the many actors required to make important decisions regarding the technology. There is a need to further elucidate the conceptual frameworks, not as two distinct conceptual framings, but as one hybrid framework which ringfences a decisional capacity consisting of beneficial risk mitigation measures but which also presenting new risks. Overall, there is a need to consider the intersection of the two main framing *moments*: that of the safety argument, which largely stresses the upside in term of risk management; and ethical critiques which underscore the downside in challenging AI moral agency and capacity to make life and death decisions.

Part Two: Conceptual Framing: The Challenges of AI Decisions

Conceptual framing is a process of using established concepts and their meanings to construct a framework around phenomenon. The construction and use of concepts creates a model of conceptual relations which bring meaning to the phenomenon. In terms of emerging technologies, the way in which the new phenomenon is conceptually framed has significant impact on how others engage with and understand the technology. This is particularly important in terms of how different actors understand technological relations, to society, and to actors who need to make decisions regarding societal engagement. As such, conceptual frameworks are key to how we understand the benefits and risks which the new technological phenomenon presents. A conceptual framework aims to configure the relations in the technological phenomenon, but this is not always a straightforward process. The conceptual framework is a way of modelling and communicating what a technology means during a phase of technological emergence. A seminal example of how a seemingly straightforward and intuitive instance of conceptual framing led to distortion and difficulties around precise meaning, resides with the concept of artificial intelligence. In his Dartmouth proposal of 1955, John McCarthy introduced artificial intelligence (AI) as the inaugural, conceptual framing of machines which carry out intelligent tasks (McCarthy 1955, 1956). Since then, the conceptualisation of AI has posed significant difficulties (Searle 1980; Kaplan, 2016; Dreyfus, 1965, 1972; Dreyfus, 1986) in terms of how different agents and actors engage with its conceptualisation (Johnson & Verdicchio, 2017). Such challenges have important ramifications with respect to the anticipatory research and anticipatory governance of the potential impacts of emerging technologies (Donk, 2012; Gasser and Almeida, 2017).

Conceptual frameworks are utilised in most areas of research that rely on clear models of relational meaning pertaining to complex entities and values; from finance (Berger and Udell, 2006), and education and models of pedagogy (Savage and Sterry, 1990), to societal impacts of technology and risk (Bachman et al., 2015; Kasperson et al. 1988). An important commonality throughout the use of conceptual frameworks concerns the attempt to control meaning; whether to improve it, to supplant established frameworks, or to create new ones. Thus, frameworks concern the ability to conceptually construct meaning for the purposes of knowledge and dissemination. Conceptual frameworks are often rehearsed as a means of investigating and reinforcing particular models obtaining to understanding and communicating the meaning of technological phenomenon. The utility of conceptual frameworks as an elucidatory exercise has proven beneficial in modelling and communicating new technologies, meaning, and societal relations (McGrath and Hollingshead, 1994). Two key qualities are intrinsic to effective and accurate conceptual framing: first, the ability to clearly present the relationship between entities or phenomena which reside together; and second, the ability to communicate these relationships to stakeholders. As conceptual framing is an essential

piece of the cognitive jigsaw and an important mechanism to convey both meaning and understanding of phenomenon (Sorokin, 2016), it is intrinsic to understanding and communicating innovative technology (Maxwell 2013, 41). At the same time, it can exert a normative impact on debates within society. Accurately anticipating the societal, ethical, and legal (SEL) impacts of emerging technologies is a process of investigation that is, in and of itself, contingent upon the initial conceptual framing of the technologies (Donk et al. 2011). Conceptual frameworks can contribute to a more informed context of conceptual meaning which determines other downstream frameworks which, in turn, determine conceptual accuracy, clarity, and transparency, such as governance frameworks (Renn 2008). The conceptual framing often begins as an ad hoc process with little consideration of the accuracy of the concepts used, and which in fact, are sometimes are arrogated from other domains (Cook and Bakker 2012).

Framing Artificial Intelligence and Autonomous Decisions

The new and emerging technological paradigm of AV has generated some technological disorientation, more specifically in respect of the decisional capacity of embodied AI products. There is a progression of conceptual meaning and conceptual framing that begins with the research phase and culminates with how the media and society engage with the concepts relating to the technology. However, given that development of innovation depends upon the key metrics of governance, the media, and public perception, there is a need for closer scrutiny of how initial framing plays out in the public arena. The literature on risk amplification speaks to this issue and points to the need for debates which set a positive and inclusive tone (Pidgeon *et al*., 2003). This is true both of the more general phenomena of risk amplification, as well as more discrete phenomena, such as dread risk. Risk amplification and fear of new and emerging technology is well-documented in the literature and suggests the care needed around initial conceptual framing (Frewer *et al*., 2002). This aspect is taken up by Johnson & Verdicchio (2017) who maintain the need to "*argue for a reframing of AI discourse that avoids the pitfalls of confusion about autonomy and instead frames AI research as what it is: the design of computational artefacts that are able to achieve a goal without having their course of action fully specified by a human programmer*" (Johnson & Verdicchio, 2017). While their critical approach bears on the challenges of framing embodied AI products such as AVs, they represent a minority who address the question. We contend that in addition to autonomy there are further related complex challenges specific to the framing of AI and AI decisionality. In relation to how effective ontological domains are set out between concepts (Franklin & Ferkin, 2006), there is a similar need to anticipate conceptual challenges in the initial framing and ontologies of concepts used in framing embodied AI products (Cunneen et al. 2018). This is essentially a call for temporal considerations to be captured in the concepts employed, as since this field is highly

dynamic, and in terms of the configuration of actors and their anticipated roles, these are liable to change over time.

The safety argument and the ethical challenge both relate to anticipating the SEL impacts of AVs regarding decisional capacity. A critical analysis of both examples suggests that the disparity arises from a failure to engage at the necessary meta-level or construct informed accurate conceptual frameworks of AV decisional capacity, and failure to consider the important differences between how society and users understand human and machine decision-making in more detail. In fact, the core question of the SEL impact of AVs is yoked to the meaning framework of machine driving decisions and human driving decisions. This underlines the necessity to interrogate the conceptual framing of AV driving decisions. Without accurate SEL impact analysis, the challenges of uncertainty and risk will hinder informed research, development, and societal perception (Renn 2008: xv). And without accurate metrics of the SEL impact, systems of governance cannot provide the mechanisms which balance the need to support innovation with the duty to assess potential risks and protect society from harms. This is particularly emphatic in cases where innovation is maintained to be ethically justifiable. In short, all innovation warrants a process of analysis by which to accurately frame the legal and general principles of associated societal rights to safety, freedom, equality, privacy, and welfare. While both the safety argument and the ethical challenge are in agreement in framing AVs to centre on the decisional capacity of vehicular driving intelligence, they offer very different matrices of the range of decisions AI must carry out to safely traverse the human road network. Each interpretation begins with the focus on decisions but frames the decision capacity differently, and each anticipates very different accounts of the potential SEL impacts of AV decisions and governance. Diverse perspectives and interpretations are an integral aspect of developing research and knowledge contexts, but as multiple agents and actors engage with the different frameworks, the potential for inaccurate framing feeding into systems of governance is a significant concern. We have two very different accounts of decisional capacity regarding the anticipation of SEL impacts and governance of AVs. Each one frames the decisional capacity in dramatically opposing ways: one claims it is a superior driving decision capacity that will save lives; the other insist it presents a risk of limited decisional capacity which could inadvertently pose significant ethical problems (Lin, 2016).[1] Proper analysis clarifies the AV decision domain, and if we are to judge by the two principal framing values of the safety argument and ethical challenge, the AV decisional framework presents a technological medium that remains conceptually obscure.

Part Three: The Safety Argument

Framing the Space regarding the Societal Benefits of Autonomous Driving Decisions

The first conceptual framing of AV, referred to as "the safety argument", expounds on the decisional phenomenon of AV to offer superior driving decisions and decisional capacity compared to human drivers. As such it examines AV decisions in terms of the safety and risk mitigation benefits inherent in the technology. However, the need to counter any potential misunderstanding that may arise remains evident. AVs are widely regarded as technologies which can outperform human driving capabilities because although humans are statistically proficient drivers and "*It is possible to drive a private car 13000km a year for fifty years with more than a 99 percent chance of survival*" (von Suntum, 1984: 160), human drivers make human errors. In fact, the NHTSA report that "recognition errors, decision errors, performance errors, and non-performance errors" contribute to 94% of the "critical reasons" of road traffic accidents (Singh 2015). Their analysis is supported by research undertaken by Evans (1996) which also concluded that "*The road user was identified as a sole or contributing factor in 94% of crashes in the US study and in 95% of crashes in the UK study*". The central claim which supports the ethical justification of the safety argument and the development and use of AV then, is that an AV which functions at least as efficiently as human drivers will eliminate many of the human errors which directly contribute to road traffic fatalities. Since this contention is far more complex than it first appears however, this paper aims to properly elucidate the reasoning of the argument.

As stated above, statistical driving data to date highlights an alarming correlation between erroneous human driving decisions and road accident fatalities. In light of this, advocates purport that AVs offer the opportunity to dramatically reduce human decision-related driving fatalities (General Motors 2018; LaFrance 2015; Bertoncello and Wee 2015; Litman, 2017). Many other commentators are equally enthusiastic about the potential safety benefits but caution that these remain purely hypothetical unless issues such as policy (Campbell 2018) and practical implementation (Grush and Niles 2018: Schiller et al. 2018) are addressed. Nonetheless, numerous states, including the U.S., U.K., and Germany, have already fast-tracked policy to support AV research and testing, and both the automotive industry (GM: 2018) and private funded research (Anderson and Kalra et al. 2014) insist that AVs afford significant safety benefits that will save countless lives. However, Campbell (2018) points out that while there may be acceptance of AV safety arguments in principle, there is prevailing confusion as to how the technology is to be supported and such benefits actualised.

The opportunity to save lives provides a compelling ethical basis to support any innovative technology, and when directly contrasted to AVs, human driving decisionality is characterised by bad or inferior decisions which cost lives. The safety argument of AV technologies stresses the core safety benefits which are built upon more accurate driving ability and supported by advanced decisional capacity to navigate the road network. The focus on decisional intelligence and capacity as a means to reduce RTAs and fatalities, ringfences the problem space: namely, that the use of AVs will avoid most types of RTAs by eliminating the opportunity for human driving decision error. It therefore follows that the success of the technology hangs on the AV capacity to make effective driving decisions and reduce the frequency of RTA loss of life and limb. The crux of the argument is that the replacement of human driver by AVs will save lives by decreasing the occurrence of distraction, intoxication, or fatigue-related RTAs. What is most interesting about the safety argument is its reliance on the technological realisation of high-level operational success. Thus for the safety argument to achieve its aims the technology must accomplish a level of effective driving decisional capacity equal to the success metrics of human driving decisional capacity. The AV ability to do this rests on their ability and decisional capacity to make more consistently accurate driving decisions than humans and traverse the human road network with at least an equal statistical record of safety performance to human drivers.

The challenge is that for AVs to accomplish the claim of reducing the frequency of human driving decision error in the core categories which make up 94% of driving errors, AVs must accomplish the global feat of improving across the spectrum of all driving decisions. There are two parts to this analysis of human driving decisions: the erroneous (bad) decisions that make up the NHTSA's 94% claim; and the global frequency of successful (good) decisions. Yet, the argument automatically defaults to a concentration on technology by removing the array of bad driving decisions which arise from key example areas. However, for AV technologies to reduce the 94% frequency of bad human driving decisions it must first achieve the operational milestone of at least equalling good human driving decisions which depends on a successful roll-out of the technology. While statistically such a benefit can further support the overall anticipated global performance metric, the primary challenge is still to reach a level of successful driving decisions equivalent to good human driving decisions. This is largely omitted from the ethical justification of safety arguments which hinge on statistical evidence of the AV capacity to outperform humans in the single category of human driving decisions. By doing so, it yokes the safety argument to decisional benefits, which, in turn, depend upon the global decision performance of the technology to outperform the most successful aspects of human decisionality.The WHO (2018) has estimated 1.35 million RTA related deaths

in 2016 and this is a figure that continues to climb. When considered in relation to the NHTSA assessment it is not surprising that many of these fatalaties relate to erroneous human driving decisions. However, the ratio of good driving decisions is simply not considered. In fact, the target figure required to justify the technology is not the ratio of decisions which account for RTA fatalities, but rather the ratio of human miles driven safely. Thus, the many points of tension within the safety argument include: (1) paradoxically, and as is the case with all forms of data-dependent analysis, human bias can compromise data accuracy and the conclusions drawn from analysis; (2), when reconsidered in light of the above critique of good human driving decisions, the claims of the safety argument will take considerable time to justify AV use.

The following section argues that the safety argument as a key metric in developing a conceptual framework of anticipatory SEL impact of AV goes too far in claiming decisional superiority while so many unresolved challenges remain, and criticisms persist regarding the efficacy, ability, and scalability of the technology.

The Problem with the Safety Argument

The safety argument has framed the safety claims of AV decisions to rest on a conception of decisionality which only addresses the most problematic decisions of human driving. It is indisputable that AV will offer a driving decision spectrum which will preclude intoxication, distraction, fatigue, and poor behavioural decisions such as speeding. However, this argument is more problematic than it appears as it depends on the overall ability of the technology to outperform the full range of human driving decisions. Indeed, whether this is achievable remains a matter of conjecture, particularly considering that AV will bring new accident metrics to the safety figures. There will undoubtedly be RTAs uniquely tied to AV decisionality in the global context, such as sensor error, programming bugs, unanticipated objects, classification error, and hardware faults. While it is hoped these will precipitate less damaging incidents due to speed limitations and safety mechanisms, they nonetheless present further challenges to AV safety arguments and the target of global improvement in driving decisions. Effectively, this means that if, for example, the key categories of human driving error could be removed from the equation of road safety statistics, the resulting figures would point to a human driving capacity, perhaps beyond that of any emerging driving technologies. The analysis claims that the premises of the safety argument and the statistical RTA figures which support the claim that AV decisions are safer than human driving decisions, cannot, at present, be maintained. The safety argument concerns the most problematic decisions of human driving, yet completely elides the immensely efficient spectrum of

decisions which human driving represents. To boost statistical gain, it builds on the removal of problematic decisions from the driving spectrum. What remains unclear, however, is that in doing so, it must provide a core decisions spectrum which is as successful as the normal majority of human driving. The latter point is the most contentious part of the argument, given that the former claim depends on the success of the latter, which will, in turn, remain unverifiable for some time. It is evident that the safety argument fails to provide an accurate account of AV decisional capacity. For AV to resolve this aspect they must achieve a consistent level of competency across the entire spectrum of human driving decisions.

The Safety Argument Does Not Define the SEL Impact of AV Decisions

Given that studies such as Blanco et al. (2016) conclude that contrasting AV driving data to human driving data cannot effectively determine the outcome, the safety argument cannot be justified by data analysis alone. Moreover, while many studies have produced figures quantifying human driving safety there are numerous issues inherent in using such data (ibid). A number of studies and sets of figures stand out in this regard. Early analysis by von Suntum maintains that "*It is possible to drive a private car 13000km a year for fifty years with more than a 99 percent chance of survival*" (von Suntum, 1984, 160).[2] More recent analysis by Hevelke & Rumelin (2015) concludes that in the period 2005 to 2009, there was one accident for every 1.46 million kilometres travelled. On this basis Goodall (2014a) calculates that "*with a Poisson distribution and national mileage and crash estimates, an automated vehicle would need to drive 725,000 mi on representative roadways without incident and without human assistance to say with 99% confidence that they crashed less frequently than vehicles with human drivers*"(Goodall 2014a). While driving statistics and analyses are an important dimension of traffic management and safety, a report by the Virginia Tech Transportation Institute questions the practice of using such figures as definitive support for safety analysis and highlights the problems inherent in relying on such data (Blanco et al. 2016). They assert that not only is the human driving data questionable due to the omission of up to 59.7% of unreported accidents, but also because the format and criteria of driverless vehicle crash reporting data presents a very different challenge (ibid). At most, it can offer one aspect of a multifaceted investigation into the SEL impact. Difficulties regarding the efficacy of contrasting driving data are also brought to the fore by Schoettle & Sivak (2015) who draw attention to the 1.5 million autonomous driving miles and the 3 trillion conventional vehicle miles travelled annually in the U.S. alone. Such contrasting figures cannot deliver the data required to develop an accurate SEL impact or support the claims of the safety argument. In fact, each research group identifies significant differences which influence the driving data; so much so,

the task has been likened to comparing "*apples and oranges*" (Blanco et al. 2016). For example, AV driving miles is predominantly defined by open road good weather driving which does not include driving in fog, adverse weather events, or snow (Schoettle & Sivak, 2015). Research groups also concur that current figures suggesting a greater frequency of crashes in autonomous driving than human driving are probably based on inaccurate findings due to the "*limited exposure of the self-driving car project to real-world driving increases statistical uncertainty in its crash rate. That uncertainty will decrease as it receives more on road in traffic testing*" (Blanco et al. 2016).

Clearly data deficiencies will persist for some time, but even when all the data is in, certain issues concerning the accuracy of contrasting human driving data to autonomous driving data will inevitably remain. For instance, questions regarding environmental variables such as weather and object classification, and which kinds of accidents qualify, will continue to be raised (Schoettle & Sivak, 2015; Blanco et al. 2016, 40). If the safety argument or similar claims that AVs will demonstrate safer driving and save lives are to be verifiable, a more complex analysis of the driving capacity, specifically regarding driving decisions, must be provided. Yet it is clear that a direct contrast between human and autonomous driving data cannot provide definitive support to claims relating to the SEL impact or otherwise. Data analysis may well support anticipatory scrutiny of the SEL impact, but it cannot be definitive, since numerous issues of data accuracy and methodologies compromise such dependencies and contest such claims. As such, it is possible that statistical data may not provide the safety argument with the justification it seeks. On the contrary, since it is a purely hypothetical argument based on the belief that technology can eliminate the key categories of human driving decisional error, the matter of whether advanced AVs can improve the safety statistics for human travel by road may remain uncertain for some considerable time. In effect, AVs must achieve 1.7 million accident-free, diverse weather, driving miles before figures can properly evaluate their performance and verify the claims of the safety argument. Considering the above criticisms, the safety argument cannot define the meaning of AV decisionality. Nor can it claim to represent the sum of AV decisions. At most, it can inform a partial understanding of one layer of the AV decisional ontology. The ethical challenge builds its critique of the safety argument on such inferences.

Part Four: The Ethical Challenge

The above safety argument asserts a framing of AV decisions that mitigate the many human driving decision frailties which constitute risk scenarios within the driving phenomenon. However, the same AV driving intelligence technologies which support and determine its decisional capacity also present

when compared to human driving decision capacity in terms of decisional limitations. In fact, AVs face difficult challenges in responding to driving events that involve the multi-layered space obtaining human values, ethics, and laws. For example, there are immense complexities inherent in programming autonomous intelligent machines to identify, process, and carry out decisions which conform to human values and ethics. This emphasis on the ethical component of driving decisions is highlighted by Gerdes & Thompson (2016) who suggest that the decisional capacity of AVs will ultimately be judged not by statistics or test-track performance, but rather through the "ethical lens" of the society in which such vehicles operate. They further claim that the questions of machine morality both as a capacity and as moral intelligence that supports moral analysis and decisionality, along with society's moral analysis of the AVs decisions, will determine how society anticipates the SEL impact of AVs decisions.

The second conceptual framework explored in this paper concerns possible AV decisional limitations. While the safety argument focuses on key decision benefits and risk mitigation, many commentators have identified the potentially obverse decisional limitations to the technology which may engender new forms of driving risks. Since it is unlikely that AV will have the capacity to make decisions which encompass human values, rights, societal norms, and ethics, one such framing relates to the "ethical limitations" of the technology. Many view this distinction as a significant technological deficit which, in the case of unavoidable road traffic accidents, could present new risks to society and users. For example, certain programmable AV decisions will intrinsically align with predefined human rules, such driving laws, driving codes, social norms, and accepted conduct. Other decisions will be autonomous as AVs present a complex decision ontology (Cunneen et al. 2018). For AV to function they will necessarily require a wide scope of decisionality, given that that certain decisions are devised to override erroneous human instructions, and some decisions may even break laws as we know them. AV driving intelligence will consist of diverse intelligence components designed to best support navigation of the human road network. The ability of a machine to make decisions and traverse the human road network without not only causing harm, but in such a way as to be safer than human drivers, represents a new and important development in the phenomenon of human/machine relations in that it transposes the uniquely human ability to safely navigate through the world to a machine. This transfer of ability is underpinned by the claim that AVs will diminish driving risks and provide a safer driving experience. Moreover, the ability of machines to replace human drivers in this way marks an important step towards reliance on AI and the beginning of a risk-mitigation relationship wherein society increasingly looks to machines to reduce real world risks to humans. In essence, this

means that more of the risk phenomenon is allayed by transferring the risk mitigation decisional context to AI.

Patrick Lin (2011) criticises the safety argument and proposes an ethical challenge. He backs his position by utilising psychological, moral experiments, and edge cases, to provision the claim that even the most advanced AV intelligence will have limited decisional capacity. This is evident when the AV is confronted with scenarios that necessitate decision responses relating to human rights and morals (Lin 2013; 2016). Lin maintains that in the event of accidents, AVs could fail to respond to scenarios encompassing human values, through an inability to identify the metrics of the human values, ethical relations, or legal consequences of an action or inaction. As such it questionable whether, even with safety improvements in mind, manoeuvring intelligence alone can support the telos or goal of safe driving. Lin's approach is informed by research carried out by Wallach & Allen (2008) on the possibility of moral machines. These authors appeal to the trolley dilemma (TD) and argue that machine morality will be required for a "robot car" to adequately respond to driving events and specifically to unavoidable RTAs. Lin also appeals to numerous hypothetical scenarios to defend his conclusion that the diversity of the human environment and road network will require driverless technology to have moral intelligence:

> *If motor vehicles are to be truly autonomous and able to operate responsibly on our roads, they will need to replicate-or do better than-the human decision-making process. But some decisions are more than just a mechanical application of traffic laws and plotting a safe path. They seem to require a sense of ethics, and this is a notoriously difficult capability to reduce into algorithms for a computer to follow.*
>
> *(Lin, 2015)*

As variations in "the trolley dilemma", edge cases are generally used to evaluate emotive scenarios and choices regarding the lesser of two evils. When confronted with a no-win scenario, the AI of a driverless car must an immediate response to an unavoidable RTA. Such a dilemma challenges the AI to make a decision that will inevitably lead to the death of at least one person. Complex, no-win, lesser-of-two- evils type scenarios are therefore simulated using emotive values such as a parent and child, elderly women, children and school buses, or even a doctor, to interrogate social norms, moral expectations, and individual reasoning.[3] The complex relational interplay which frame such moral dilemmas arguably support the view that moral intelligence is prerequisite for AVs to make accurate, informed decisions in response to life-threatening, but not, inescapable, driving events (Lin 2015). In short, AVs will require the capacity to engage with real-life human dilemmas in order to carry out the function of safely traversing the human road network.

The Ethical Challenge

Premise 1: The human road network is laden with human values

> *The human road network encompasses numerous unpredictable scenarios consisting of variations of human agents in the form of human drivers, pedestrians, cyclists, children, and animals. This environment will force morally loaded events and scenarios upon the driving intelligence of an automated system.*

Premise 2: AV will necessarily make decisions which involve human values

> *When confronted with events and scenarios which involve human values, possible harm, or loss of life, AVs will make decisions which directly impact on human welfare, even if they are only programmed to classify objects they perceive in the environment, according to classifiers of size and shape.*

Conclusion:

> *Our human road network presents a diverse and fluid environment. All vehicles negotiating this environment will undoubtedly be confronted with unexpected events, potential collisions, and life or death RTAs. AV responses to such moral scenarios depend on the intelligence frameworks which determine its actions. In the absence of moral intelligence AVs will mediate moral scenarios as non-moral scenarios via a value spectrum which is solely predicated on relational quantifications between the individuals and objects.*

AV will respond and make decisions based on limited data and hold to identifiable values relating to object metrics of size, mass, and speed. Lin's application evinces the confusion as one that seeks to address human driving intelligence (HDI) and its unique contextual meaning to be isomorphic to artificial driving intelligence (ADI) (Cunneen et al. 2018). Lin's use of the trolley dilemma applies a human moral and psychological thought experiment to ADI. While such a scenario can apparently support doing so, the notion of confronting an ADI with such a decision is problematic for several reasons. The challenge is in understanding why. It is only by considering the ontological differences in HDI and ADI as driving intelligence supporting determined decision capacities and spectrums of decisions, that the significant differences between them become clear. Nonetheless, the emotive strength of the scenario has been integral to the dominance of many negative news headlines. It has also played a role in detracting from questions about the compatibility of human moral experiments to AI and autonomous machines. Lin's approach is supported by similar analysis developed by Noah Goodall (2014a), who presents "*three arguments: that even perfectly functioning automated vehicles will crash, that certain crashes require the vehicle to make complex ethical decisions and that there is no obvious way to encode human ethics in computers. Finally, an incremental, hybrid approach to develop ethical automated vehicles is proposed*" (Goodall 2014a). Goodall appeals to the two further predicates of Lin's argument: he maintains that AVs will themselves crash; he purports that since unavoidable RTAs will present

scenarios obtaining to human values, AVs need moral intelligence. As Goodall concludes, the main problem is the absence of current programming technology which can support the ethical resolution of complex moral decisions within the AV decisional spectrum. In other to do this, driving intelligence would need to be augmented with moral intelligence in order to process identifiable moral dilemma data and generate moral decision options within the decisional spectrum. Goodall's approach is particularly interesting in that he responds to this difficulty by postulating options which could support moral decision processing. He is not alone in this.

Many others have also claimed that moral theorisations can be programmed into machine intelligence.[4] Goodall's view supports the ethical challenge by claiming that; *"If, however, injury cannot be avoided, the automated vehicle must decide how best to crash. This decision quickly becomes a moral one…" (*Goodall 2014a, *60)* This contention is echoed by Korb (2007) who claims that all AI inhere a necessary duty to ensure that the intelligence has cognisance of its actions being adjudged to be right or wrong. The AI must have parameters of acceptable, unacceptable, right and wrong behaviour if the autonomous behaviour is achievable.

> *But if we build a genuine, autonomous AI, we arguably will have to have built an artificial moral agent, an agent capable of both ethical and unethical behaviour. The possibility of one of our artefacts behaving unethically raises moral problems for their development that no other technology can.*
>
> *(Korb, 2007)*

The possibility of programming machines to make ethical decisions is highly controversial. For instance, ensuring that the processing of ethical datasets does not contain programming biases raises concerns. Verifying that the decisions carried out are lawful and embody our individual and societal values also pose difficulties, as do issues of transparency of manner and access. Such challenges suggest programming ethics into socially embedded machines to be an insurmountable task: all the more so when they relate to decisions which have a direct bearing on human harm or loss of life. There are two strands of response to this situation. The first represents the general view of the industry and attests to insignificant applications in respect of the issues surrounding moral decisionality in autonomous vehicles. As such, the practical application of moral decision intelligence is not deemed significant enough to warrant the research, investment, and/or development of a moral decisional capacity for autonomous vehicles. This cost/benefit view of moral decision capacity is refuted by technological ethicists such as Lin, who argue that no matter how low the frequency of moral dilemmas confronting autonomous vehicles, it should be intrinsic to the decision capacity of autonomous vehicle programming. Lin's critiques have fuelled research attempts to address the challenge of programming ethics into autonomous vehicles.

The move to appeal to an artificial moral agent is often contextualised in emotive life or death scenarios of decisions or actions taken by a vehicles operating system about unavoidable RTAs. However, Goodall raises the important point that certain crashes are unavoidable. Thus, no matter how good the sensory and impact avoidance software is, there should always be some calculation for damage limitation when AVs will crash. Goodall follows the standard approach to AVs actions about some "inevitable collision states" (Fraichard and Asama, 2004), as it is evident that the standard approaches to AVs and RTAs frame actions as moral actions. Goodall maintains that "an automated vehicle's decisions that preceded certain crashes had a moral component", and the challenge concerns the reality that "there was no obvious way to encode complex human morals in software" (Goodall, 2014a, 58). A similar view is put forward by Kumfer et al. (2016) in "*A Human Factors Perspective on Ethical Concerns of Vehicle Automation*". This paper stresses that nothing in current AV programming supports moral decisionality and perhaps it never will. This creates a scenario where we are assessing the moral actions of AVs, but the only moral context is the human moral framework we are applying to it. Once again, this is an unexpected double-bind requires further investigation. Goodall's emphasis on optimal crash strategies demonstrates how the pre-crash window presents an example of decisional limitations, and how such an ability introduces morality into the decisional capacity by supporting the need to investigate how moral values might be applied (Goodall, 2014a). This directly addresses what Goodall denotes as a clear deficiency in governance and U.S. state law to cater for "computerized control of pre-crash or crash avoidance behaviour" (ibid). Goodall is undoubtedly correct in his assessment of the need to investigate such behaviour, especially as it relates to what he describes as crashes with a moral component. However, Goodall, like many others who subscribe to this standard interpretation of approaching the operating system governing control of AVs as artificial moral agents, has unnecessarily and emotively weighted the analysis. As soon as a hypothesis of artificial agency is applied to a device with interactive qualities, it expands to a far more complex conceptualisation of artificial moral agency. As such, there is not only a need to clarify the conception of artificial agency, but also, to address the seemingly unavoidable expansion to the artificial moral agency and to elucidate and qualify the application of both conceptions of agency.

> *…an AV cannot cope with situations that were not anticipated and taken into account by the programmer (even if it includes a learning approach). Overall, an AV's decision making is imperfect and uncertain.*
>
> *(Dogan et al, 2016)*

The Ethical Challenge

Ethical tensions align with the various ways decisions carried out by AVs reflect moral preferences as general principles, such as to ensure our safety, and cause no harm to others. For this reason, moral challenges in the form of possible decisions are bound to confront the technology. Lin is undoubtedly accurate in this respect (ibid). Nonetheless, AV morality may not comparable to human morality, and therefore unamenable to assessment by the same metrics of morality we use to judge people. This is one reason why the focus on decision intelligence and capacity is essential, but it should also be framed in such a way that can be appropriately conceptualised without anthropomorphising the decisional process. While this tendency is more prevalent outside the academy, anthropomorphises may nonetheless impact the media discourse which plays an important role in formulating public opinion and policy. Analogous debates around genetically modified crops are instructive in this regard (Frewer *et al.*, 2002; Turney, 1998).[5] Our case here is that in order to address this tendency for a certain type of Rhizome-like framing (Deleuze and Guattari, 2009) which takes the discourse in a certain direction it is necessary to revert to the conceptual frame. This is all the more so when debates brush against mythic narratives which anthropomorphise the issue. To accurately understand the moral context of AV decisionality then, numerous layers of analysis must be ontologically elucidated (Cunneen et al. 2018). Questions such as how the machines were programmed to respond to a given scenario, how they classified key objects, people and relationships, and whether the moral analysis was predetermined or arose from accurate autonomous moral decisions as a result of machine learning or an adaptive algorithm, must be addressed.

Yet, the unreliable advocacy of the safety arguments which compare AV and human decision data based on statistical data analysis, along with the biases of public perception, arguably impede any accurate anticipation of the SEL impact of AV. The field of ethics, and particularly applied ethics, boasts a tradition of vanguarding societal safety by investigating tensions when law and policy fail to present the issue in a format which satisfies all interested parties, such as nuclear fuel, nuclear arms, and child labour. However, in the context of AV, even ethics has been criticised for its apparent failure to accurately anticipate the SEL impact of AV. Considering recent ethical critiques of the safety argument, many commentators question the format and use of ethical challenges in focusing on the question of moral limitations of AV decisional capacity, while Nyholm & Smids (2016) maintain the ethical challenge developed by Lin (2011, 2012, 2015) is confused. Their analysis is centred on Lin's use of variations of the TD to support his hypothesis (Lin, 2011, 2015) which underscores the clear disanalogies between AVs and the use of the TD (Sven and Smids, 2016). Noah Goodall, a vocal supporter of Lin's ethical challenge to AV, also flags the difficulties posed by the use and misuse of the TD in elucidating AV decision limitations (Goodall,

2013). Charisi et al. (2017) similarly criticised the TD as a means of supporting analysis of the SEL impact, maintaining that the question should rather be focused on ensuring that the ethical programming of the AV decisional capacity is achievable in a transparent manner (ibid). Lin himself has also come to acknowledge the contingent nature of the TD in the context of AV, but insists that as it cannot be completely dismissed as a potential scenario wherein AV will be confronted with weighing moral decisions, the challenge persists (Lin, 2015).

An Ethical Lens: Societal Perception and Anticipating SEL Impact

Lin's criticisms of the safety argument have important ramifications in relation to the public use of the technology, public risk perception, and informed consent. If we are to judge based on media headlines such as "killer cars", the challenge of machine morality has already coloured public perception and risk perception of an ability of AV to make autonomous moral decisions.[6] If Gerdes and Thornton (2016) hypothesis is proved correct, society's perception of AVs will be the key determinant of the SEL impact of the technology. Their stance privileges social perception and conceives judgement of autonomous technologies to be bound by society's "ethical lens" (Gerdes and Thornton, 2016). Public risk perception is clearly averse to AV accidents that are highly publicised, even when countered by claims that millions of miles of safe driving have been achieved. However, when one considers the numerous news headlines and the data obtained from public questionnaires (Bonnefon *et al.*, 2016) regarding AV safety and user perception, societal concerns align with the ethical challenge and contrast against the claims of the safety argument.

The importance of understanding the concept of AV decisionality to anticipate the SEL impact of the technology has been closely discussed and defended in this paper, despite evidence that both research and public perception already misunderstand it. As such, there is a clear need to provide an accurate account of autonomous decisional capacity by elucidating the concept of AV decisionality. This will require ring-fencing the limitations of decisional capacity as the technology evolves. This contrasts both positions and takes the view that the crux of the SEL impact will concern the decisions that AV can or cannot make. This disparity in interpreting AV decisional capacity identifies an underlying difficulty in terms of framing AV and autonomous technology decision-making. Autonomous vehicles present an autonomous decision-making agency immersed in one of the most fast-paced and unpredictable aspects of modern society; namely, the human road network. This phenomenon of driving is known to elicit emotive responses from the most rational of people. As such, it inhabits an unpredictable emotive human space that AVs must quantify in numerous different ways. The

top-tier challenge therefore is to produce a technology which can satisfy the demands and expectations of a public that will, many suggest, have a low tolerance of AV decisional errors. Carl Nash (2017), for instance, contends that low public tolerance for AV-related fatalities could have unexpected effects on the stakeholders involved in developing the technology.[7] The ultimate success of AV then, will depend upon a combination of factors contributing to general social perception, comprising society's ethical lens (Gerdes & Thompson, 2016) and society's risk perception (Litman, 2017) of the technology. If this is an accurate assessment of the underlying drives which will ultimately determine the SEL impact, then the focus on AV decisionality will effectively be instrumental to and anticipate it.

Conclusion

The autonomous vehicle is but one application of artificial intelligence technologies which have an enormous bearing on contemporary and future society. As AI implementations become more diverse, and indeed ubiquitous, there will a greater need to understand the different contexts of decisional application. Essentially this means that in order to accurately frame each unique decision context the technology must to be taken at face value and a nonlinear relational model of classification concepts created. The implication is that we should not strive for a general "one size fits all" conception of artificial intelligence application as no single framework can sufficiently account for the numerous possibilities of machine decisional applications. Conceptual framing is an integral part of an epistemological continuum which supports governance and regulation. Given the strategic position is occupies in chains of meaning which promote public debate, it is vital that such framing is open to interrogation. Overall, this paper examines how the crucial process of conceptual framing has received relatively little attention from academics in the area of anticipatory governance. The central argument is that conceptual framing has downstream effects in terms of the debates on the governance of automated vehicles. It is a nuanced argument in that the idea of the unitary conceptual framework is not posited per se. Rather an acceptance of the limitations of current debates is acknowledged, along with calls for more precision in the construction of such frameworks, and perhaps paradoxically, an exploration of the value of reflexive approaches to conceptual framing. In fact, it is inherently self-defeating to insist on complete and final conceptual frameworks. Instead, conceptual framing should be a more reflexive practice with an iterative component, as it is not just a matter of accuracy in terms of the concepts used, but rather a realisation of the impact of such framing that counts. Since both the safety arguments and the debates around the ethics of AVs launch discourse in a particular direction, there is a need to revisit the initial framing and to adopt a more pluralistic outlook in

terms of that conceptual framing. Such is the complexity and amplification of debates around the societal impact of AVs that there is a tendency to pare back the debate for more general consumption. While this is normal in any field, is particularly prevalent in the area of emerging technology. For instance, debates on the risks posed by nanotechnology demonstrate similar weaknesses in terms of initial conceptual framing and the results have been regulatory shortcomings and widespread confusion amongst stakeholders. Reintroducing precision into conceptual framing, and indeed an acceptance that conceptual frameworks occupy an important position in any hermeneutic cycle, can help move debates on the deployment of AVs forward. The subsequent payoff enables technologies with positive SEL impacts to be better supported, while technologies with potentially negative SEL impacts can be framed more accurately, and through properly informed assessment, can be further developed.

Notes

1. As will be outlined in the paper there is a clear risk that inaccurate conceptual frameworks can have adverse and serious ramifications for the investment, governance, and public perception of technologies. This is immensely problematic when public perception holds to emotive and unsubstantiated claims relating to potential adverse impacts and risks. There are numerous examples from genetically modified foods, global warming to vaccinations. Each presents clear examples of a negative public response to innovation due to difficulties in how innovation is conceptually framed, and communicated.
2. For a more recent breakdown of RTA figures and related causes, see figures at: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115.
3. MIT Moral Machine experiment .
4. Interestingly, Yilmaz et al (2016) point out that research into machine ethics now affords a realistic scenario that will also contribute to a better understanding of human ethics.
5. https://www.theguardian.com/technology/2015/dec/23/the-problem-with-self-driving-cars-who-controls-the-code
https://www.theguardian.com/science/2016/jun/14/statistically-self-driving-cars-are-about-to-kill-someone-what-happens-next.
6. A concerning pattern here relates to the dramatic imagery of Phillipa Foot's Trolley Problem as an emotive image of the challenges to machine ethics; Phillipa Foot, "The Problem of Abortion and the Doctrine of the Double Effect" (1967). Chris Urmson is critical of such philosophical discourse: https://www.washingtonpost.com/news/innovations/wp/2015/12/01/googles-leader-on-self-driving-cars-downplaysthe-trolley-problem/ .
7. If AVs fail to live up to expectations or become ensnared and held back due to a rush to develop legislation to respond to identifiable safety concerns raised by public and competing industry's critiques, as in the case of autonomous trains, the prospect of AVs arriving in any developed scale could be hindered.

Funding

This work was supported by the Horizon 2020 [690772].

References

Alic, J. (1994). The dual use of technology: Concepts and policies. *Technology in Society* 16 (2):155–72. doi:10.1016/0160-791X(94)90027-2

Allenby, B. R. (2011). *Governance and technology systems: The challenge of emerging technologies*. In *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight* (pp. 3-18). Springer, Dordrecht

Anderson, James M., Nidhi Kalra Stanley, K. D., Sorensen, P., Samaras, C., & Oluwatola, O. A. (2014). *Autonomous vehicle technology: A guide for policymakers*. Santa Monica, CA: Rand Corporation

Asveld, L., and S. Roeser. 2009. *The ethics of technological risk*. London: Routledge.

Bachmann, R., N. Gillespie, and R. Priem. (2015). Repairing trust in organizations and institutions: Toward a conceptual framework. 36 (9):1123–42

Berger, A. N., and G. F. Udell. 2006. A more complete conceptual framework for SME finance. *Journal of Banking & Finance* 30 (Issue):11. doi:10.1016/j.jbankfin.2006.05.008.

Bertoncello, M., and D. Wee 2015. Ten ways autonomous driving could redefine the automotive world. Accessed: http://www.mckinsey.com/industries/automotive-and-assembly/our-insights/ten-ways-autonomous-driving-couldredefine-the-automotive-world

Blanco, M., J. Atwood, S. Russell, T. Trimble, J. McClafferty, and M. Perez. (2016). Automated vehicle crash rate comparison using naturalistic data. Virginia Tech Transportation Institute Report. http://www.vtti.vt.edu/PDFs/Automated%20Vehicle%20Crash%20Rate%20Comparison%20Using%20Naturalistic%20Data_Final%20Report_20160107.pdf (Accessed December 19, 2017).

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2015) *Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars?*. (Accessed November 2017)Available at: http://arxiv.org/pdf/1510.03346v1.pdf

Bonnefon, J. F., Shariff, A., & Rahwan, I.. 2016. The social dilemma of autonomous vehicles. *Science* 352 (6293):1573–76. doi:10.1126/science.aaf2654

Bringsjord, S., and A. Sen. 2016. On creative self-driving cars: hire the computational logicians, fast. *Applied Artificial Intelligence* 30 (8):758–86. doi:10.1080/08839514.2016.1229906

Burgess, J., and J. Chilvers. 2006. Upping the ante: A conceptual framework for designing and evaluating participatory technology assessments. *Science and Public Policy* 33 (10):713–28. doi:10.3152/147154306781778551.

Campbell, H. 2018. Who will own and have propriety over our automated future? Considering governance of ownership to maximize access, efficiency, and equity in cities. *Transportation Research Record*, 2672(7), 14-23

Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., ... & Yampolskiy, R. (2017). Towards moral autonomous systems. arXiv preprint arXiv:1703.04741

Choudury, C. (2007) '*Modelling driving decisions with latent plans*'. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge. doi: 10.1094/PDIS-91-4-0467B

Coates, J. F., and V. T. Coates. 2016. Next stages in technology assessment: Topics and tools. *Technological Forecasting & Social Change* 113:112–14. doi:10.1016/j.techfore.2016.10.039.

Coeckelbergh, M. 2016. 'Responsibility and the moral phenomenology of using self-driving cars'. *Applied Artificial Intelligence* 30 (8):748–57. doi:10.1080/08839514.2016.1229759.

Cook, C., and K. Bakker. 2012. Water security: Debating an emerging paradigm. *Global Environmental Change* 22 (1):94–102. doi:10.1016/j.gloenvcha.2011.10.011.

Cunneen, M., Mullins, M., Murphy, F., & Gaines, S. (2019). Artificial Driving Intelligence and Moral Agency: Examining the Decision Ontology of Unavoidable Road Traffic Accidents through the Prism of the Trolley Dilemma. Applied Artificial Intelligence, 33(3), 267-293

Deleuze, G., and F. Guattari. 2009. *A thousand plateaus*. Berkeley, CA: Venus Pencils.

Dogan, *et al*. (2016). Ethics in the design of automated vehicles: The AVEthics project. (accessed March 21st, 2017). Available at: http://ceur-ws.org/Vol-1668/paper2.pdf.

Donk, A., J. Metag, M. Kohring, and F. Marcinkowski 2011. Framing emerging technologies: Risk Perceptions of nanotechnology in the German press. *Science Communication* 34 (1):5–29. doi: 10.1177/1075547011417892.

Dreyfus, H. 1979. *What computers can't do*. New York: MIT Press.

Dreyfus, H. 1986. *Mind over machine: The power of human intuition and expertise in the era of the computer*. Oxford, U.K.: Blackwell

Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, and B. Schafer. 2018. AI4 People: An ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds and Machines* 28 (4):689–707. doi:10.1007/s11023-018-9482-5

Fraichard, T., and H. Asama. 2004. Inevitable collision states: A step towards safer robots? *Advanced Robotics* 18 (10):1001–24. doi:10.1163/1568553042674662

Franklin, S., & Ferkin, M. (2006). An ontology for comparative cognition: A functional approach. *Comparative Cognition & Behavior Reviews, 1*.

Frewer, L. J., S. Miles, and R. Marsh. 2002. The media and genetically modified foods: evidence in support of social amplification of risk. 22 (4):701-711.

Gasser, U., and V. A. F. Almeida. November 2017. A layered model for AI Governance. *IEEE Internet Computing* 21 (6):58–62. doi: 10.1109/MIC.2017.4180835.

General Motors. 2018. *Self-Driving Safety Report*. Detroit: General Motors.

Gogoll, J., and J. Müller. 2016. Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics* 23 (3):681–700. doi:10.1007/s11948-016-9806-x.

Goodall, N. 2014a. Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board* 2424:58–65. doi:10.3141/2424-07.

Goodall, N. J. 2014b. Machine ethics and automated vehicles. In *Road vehicle automation*, ed. S. Beiker and G. Meyer, 93–102. Switzerland: Springer.

Goodall, N. J. 2016. Away from trolley problems and toward risk management. *Applied Artificial Intelligence* 30 (8). doi:10.1080/08839514.2016.1229922.

Grush, B., and J. Niles, 2018. The end of driving: transportation systems and public policy planning for autonomous vehicles.

Hevelke, A., and J. Nida-Rümelin. 2015. Responsibility for crashes of autonomous vehicles: An ethical analysis. *Science and Engineering Ethics* 21 (3):619–30. doi:10.1007/s11948-014-9565-5.

International, S. A. E. 2016. *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. Warrendale, PA: SAE International.

Kaplan, J. (2016). Artificial Intelligence: What everyone needs to know. Oxford University Press

Johnson, D.G. & Verdicchio, M. Minds & Machines. (2017). 27: 575. https://doi.org/10.1007/s11023-017-9417-6

Kasperson, R. E. (2009). "Coping with Deep Uncertainty" in Bammer, G and Smithson M (ed.), Uncertainty and Rük: Multidisciplinary Perspectives. London: Earthscan.

Kasperson, R. E., O. Renn, P. Slovic, H. S. Brown, J. Emel, R. Goble, J. X. Kasperson, and S. Ratick. 1988. The social amplification of risk: A conceptual framework. *Risk Analysis* 8:177–87. doi:10.1111/j.1539-6924.1988.tb01168.x.

Korb, K. B. 2008. *Encyclopedia of Information Ethics and Security*. ed. M. Quigley, 279–84. Hershey PA USA: Information Science Publishing.

Kumfer, W. J., S. J. Levulis, M. D. Olson, and R. A. Burgess 2016, September. A human factors perspective on ethical concerns of vehicle automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 60, No. 1, pp. 1844–48). Sage CA: Los Angeles, CA: SAGE Publications.

Kyriakidis, M., J. de Winter, N. Stanton, T. Bellet, B. van Arem, K. Brookhuis, M. Martens, K. Bengler, J. Andersson, N. Merat, et al. 2017. A human factors perspective on automated driving. *Theoretical Issues In Ergonomics Science*, pp.1–27

LaFrance, A. Self-driving cars could save 300,000 lives per decade in America. 2015. Available at: http://www.theatlantic.com/technology/archive/2015/09/self-driving-cars-could-save300000-lives-per-decade-in-america/407956.

Lin, P. (2013). *The ethics of saving lives with autonomous cars are far murkier than you think*. Available at: http://www.wired.com/2013/07/the-surprising-ethics-of-robot-cars. (Accessed January 15, 2017).

Lin, P. (2015). *Autonomous driving: technical, legal and social aspects*. 'Implementable Ethics for Autonomous Vehicles.' Autonomes Fahren, edited by Markus Maurer, J. et al. Available at: http://link.springer.com/10.1007/978-3-662-45854-9_5_OnlinePDF. Berlin: Springer Open.

Lin, P. 2016. Why ethics matters for autonomous cars. (pp. 69-85), In *Autonomes Fahren*, ed. M. Maurer, J. Gerdes, B. Lenz, and H. Winner. Berlin, Heidelberg: Springer Vieweg

Litman, T. 2017. Autonomous vehicle implementation predictions (p. In *28)*. Victoria, Canada: Victoria Transport Policy Institute.

Litman, T. 2017. Autonomous vehicle implementation predictions; implications for transport planning. *Victoria Transport Policy Institute accessed*: 8 September 2017

Loh, W., and J. Loh. 2017. Autonomy and responsibility in hybrid systems: The example of autonomous cars. In *Robot ethics 2.0. From autonomous cars to artificial intelligence*, ed. P. Lin, K. Abney, and R. Jenkins, 35–50. New York: Oxford University Press

Malle, B. 2014. Moral competence in Robots. *Frontiers in Artificial Intelligence and Applications* 273:189–98.

Marchant, G. E. 2011. The growing gap between emerging technologies and the law. In *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem*, ed. G. E. Marchant, B. R. Allenby, and J. R. Heckert. Springer. (pp. 19-33). Springer, Dordrecht

Maxwell, J. A. 2013. *Qualitative research design: An interactive approach*. Thousand Oaks, Calif: SAGE Publications.

McCarthy, J. (1955). *Dartmouth proposal*, available at: http://wwwformal.stanford.edu/jmc/history/dartmouth/dartmouth.html

McCarthy, J. (1996). *What has AI in common with philosophy?* Available at: http://wwwformal.stanford.edu/jmc/aiphil.pdf (Accessed December 15th, 2017).

McGrath, J. E., and A. B. Hollingshead. 1994. *Groups interacting with technology: Ideas, evidence, issues, and an agenda. Sage library of social research, 194*. Thousand Oaks, CA, US: Sage Publications, Inc.

Millar, J. (2016). An Ethics Evaluation Tool for Automating Ethical Decision-Making in Robots and Self-Driving Cars, Applied Artificial Intelligence,30(8):787-809, DOI:10.1080/08839514.2016.1229919

Nash, C. (2017). Self-driving road vehicles and transportation planning, (accessed December 14th, 2017) https://www.law.gwu.edu/sites/g/files/zaxdzs2351/f/downloads/Self-Driving-Road-Vehicles-andTransportation-Planning.pdf

National Highway Traffic Safety Administration. (2016). *Federal automated vehicles policy: accelerating the next revolution in roadway safety*. US Department of Transportation.

Pidgeon, N., Kasperson, R. E., & Slovic, P. (Eds.). (2003). *The social amplification of risk*. Cambridge University Press.

Renn, O. 2008. Risk governance. London: Routledge, doi:10.4324/9781849772440

Savage, E., and L. Sterry. 1990. *A conceptual framework for technology education*. Reston, VA: International Technology Education Association.

Schiller, P. L., J. R. Kenworthy, N. Aarsæther, T. Nyseth, A. Røiseland, P. McLaverty, R. N. Abers, M. Douglass, J. Friedmann, R. N. Abers, et al. 2018. When Tesla's autopilot goes wrong. In *An introduction to sustainable transportation: policy, planning and implementation*, vol. 1, 3. 1–10, Aldershot, UK: Lutterworth Press, Cambridge, MA, and Chelsea Green Publishing Company

Schoettle, B., and M. Sivak. 2015. *A preliminary analysis of real-world crashes involving self-driving vehicles*. University of Michigan Transportation Research Institute.

Searle, J. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3:417. doi:10.1017/S0140525X00005756.

Singh, S. (2015, February). Critical reasons for crashes investigated in the national motor vehicle crash causation survey. (Traffic Safety Facts Crash•Stats. Report No. DOT HS 812 115). Washington, DC: National Highway Traffic Safety Administration.

Sven, N., and J. Smids. 2016. The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice* 19 (5):1275–89. doi:10.1007/s10677-016-9745-2.

Trappl, R. 2016. Ethical systems for self-driving cars: An introduction. *Applied Artificial Intelligence* 30 (8):745–47. doi:10.1080/08839514.2016.1229737.

Turney, J. 1998. *Frankenstein's footsteps: Science, genetics and popular culture*. New Haven, London: Yale University Press.

Von Suntum, U. 1984. Methodische probleme der volkswirt-schaftlichen bewertung von verkehrsunfallen (methodological problems connected with the economic evaluation of traffic accidents). *In: Zeitschrift Fur Verkehrswissenschaft* 55:153-167.

Wallach, W., and C. Allen. 2009. *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Weizenbaum, J. 1976. *Computer power and human reason: From judgment to calculation*. San Francisco, CA: W. H. Freeman

Wiener, G., and B. Walker Smith (2013). Automated driving: Legislative and regulatory action. [Retrieved November 14, 2016] Available from http://cyberlaw.stanford.edu

Wiener, N. 1960. Some Moral and Technical Consequences of Automation, Science. *New Series* 131 (3410):1355–58. May 6, 1960. Published by: American Association for the Advancement of Science Stable http://www.jstor.org/stable/1705998

World Health Organziation. (2018). Global Status Report on Road Safety 2018. Geneva, Switzerland.

Yilmaz, L., Franco-Watkins, A., & Kroecker, T. S. (2017). Computational models of ethical decision-making: A coherence-driven reflective equilibrium model. *Cognitive Systems Research, 46*, 61-74.

Young, S. (2016). The moral algorithm: how to set the moral compass for autonomous vehicles moral decisions by autonomous vehicles and the need for regulation. (Accessed May 9, 2017). Available at: https://gowlingwlg.com/getmedia/0eb5a71b-37fb-4ea9-a2c5-065fbc0d1e10/161205the_moral_algorithm