Algorithms of Life and Death: A Utilitarian Approach to the Ethics of Self-Driving Cars

Stephen Bennett

Bachelor of Arts, Philosophy

A thesis presented to Macquarie University in partial fulfilment for the degree of Master of Research (MRes)

Department of Philosophy

Macquarie University

October 2018

Contents

Abstract

The thought of self-driving cars operating on our roads is no longer the stuff of science fiction. Many major brands are currently developing the technology, and some states within Australia have already begun conducting self-driving vehicles trials. Maintaining a focus on Australia, we find that over one thousand people are killed in traffic accidents each year, and more than thirty-five thousand suffer injuries which require hospitalisation. Moreover, traffic accidents impose serious pressure on public resources, costing Australia's economy, for instance, around thirty billion dollars per year. As approximately ninety percent of traffic accidents are attributed to human error, self-driving cars are poised to significantly reduce traffic accidents and impact the associated consequences, given that they are anticipated to outperform humans in many of the tasks involved in operating a vehicle. However, whilst self-driving cars promise so much, their development raises some serious ethical questions. I argue for a utilitarian approach to the problem of self-driving programming, focusing on the dilemma of unavoidable accidents in which all courses of action result in someone being harmed. In doing so, I make use experimental philosophy, evolutionary psychology, traffic accident data, and consider how the implementation of self-driving cars could have significant impact beyond traffic accidents. Given what is at stake, and the speed of technological progression, this is a serious ethical issue, and one that is ready to be addressed.

Statement of Originality

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Stephen Bennett
October 2018

Acknowledgements

Sincere thanks to my supervisor Professor Neil Levy, my father Brian Bennett, and my partner Fayna Fuentes, for providing invaluable feedback and encouragement throughout the development of this project.

Introduction

In Australia alone, there are over one thousand people killed in road traffic accidents each year, and more than thirty-five thousand cases of serious injury requiring hospitalisation. This means that for every hundred-thousand people, between five and six will die in traffic accidents, with approximately half of all these being Vulnerable Road Users; that is, pedestrians, cyclists, and motorcyclists (BITRE 2017). To put this into perspective, if we take the student population of Macquarie University, we could expect up to three Macquarie students to be killed in traffic accidents every year. Strikingly, it is widely held that approximately ninety percent of all traffic accidents are the result of human error. This is often due to people operating vehicles under the influence of drugs or alcohol, aggressive driving behaviour, driving while tired, speeding, or being distracted by devices such as mobile phones (Kirkpatrick 2015: 19; Fleetwood 2017).

Beyond the significant consequences for those directly involved, not to mention the grief and hardships experienced by the family and friends of the people who are injured or killed, traffic accidents also impose a huge cost on society. To give an example of the pressure they place on medical resources, it has been reported that sixty to seventy percent of all people treated by trauma surgeons have been involved in a traffic accident. And in terms of Australia's economy, traffic accidents come with a cost of around thirty billion dollars per year, or seventy million dollars per day: a sum which is equivalent to Australia's entire annual national defence budget (Carslake 2017; RSA 2018). In light of such considerable costs, the development of some type of intervention that could reduce or eliminate traffic accidents would be worthy of serious consideration. In fact, given that many lives, a significant amount of suffering, as well as vast resources are at stake, I contend that this is a pressing ethical issue and one that falls within the realm of applied ethics.

In terms of traffic accidents, self-driving cars look to be a significant intervention, specifically, one that has great potential to save lives, mitigate harm, and increase public welfare. With the rapid development of technology, self-driving cars are no longer the stuff of science fiction films, nor vague concepts that lay somewhere far off in the future. Indeed, many major brands such as Google and BMW are currently developing the technology, with Google stating that they aim to bring self-driving cars to the consumer market by 2020 (Tam 2012). In Australia, many states are already trialling the technology, with early phase trials of

self-driving vehicles presently being conducted in both Western Australia and New South Wales (RAC 2018; TNSW 2018). So regardless of whether or not Google's aim is a little optimistic, it really is just a matter of time before self-driving cars are ready to be unleashed on the public at large.

The most notable impact that self-driving cars are set to have is their ability to greatly reduce traffic accidents. In contrast to human drivers, who are the cause of the majority of accidents, self-driving cars are anticipated to outperform humans in many of the tasks involved in operating a vehicle, such as having better reaction times, an unfaltering capacity to adhere to road rules, along with a greater ability to perceive objects. Impressively, they are likely to have the power to track different objects in multiple directions at the same time, all the while still carrying out general functions such as maintaining speed and navigating towards an intended destination safely. Moreover, self-driving cars will lack the problematic emotional and psychological traits that surely play a role in many accidents. For example, self-driving cars will never feel 'wronged' whilst driving in traffic and have an aggressive emotional response that leads to an outburst of dangerous driving. Nor will their judgements or driving abilities be affected by a lack of sleep or distractions such as mobile phones and stereo systems. Instead, self-driving cars will follow the rules of the road that they are programmed with to the letter, and their actions will be predictable to other self-driving cars, if not directly communicated inter-vehicle (Kirkpatrick 2015: 19). Consequently, self-driving cars promise to significantly improve road safety and reduce the pressure on a number of public resources, amongst other things.

That said, self-driving cars will inevitably encounter unavoidable accidents, and they will have to be programmed with algorithms that will direct their behaviour well before they are let out onto our roads (Bonnefon et al. 2015; Greene 2016; Kirkpatrick 2015). As such, we are forced to contend with a serious ethical dilemma which requires us to decide who to protect, or conversely, who to harm, in the event of unavoidable accidents in which all possible courses of action will result in someone being harmed. For example, when confronted with a wayward pedestrian, a self-driving car may be limited to two courses of action: protect its passenger, at the cost of striking the pedestrian, or, take evasive action to spare the pedestrian, though at considerable risk to, or even certain loss of, its passenger's life. It is this type of ethical dilemma that my thesis is focused on, and I intend to develop an argument that will direct the programming of self-driving cars in relation to such scenarios.

In treating self-driving car programming as a problem in applied ethics, my aim is to present a clear approach to the issue that may be useful, or even influential, for those who are engaged with the topic at a practical level, such as public policy makers and vehicle manufacturers. That said, for real-world problems such as this, it is not only necessary to present clear and compelling arguments, but such arguments need to be based on solid data and well supported theory. It is all well and good to speculate when it comes to abstract questions and hypothetical scenarios, however, when dealing with questions that involve actual harm, it would be irresponsible to make claims without reference to tangible empirical evidence, particularly if one intends for their views to have any sway amongst those whose job it is to actually turn philosophical argument into action. Consequently, I will make use of traffic accident data, findings from experimental philosophy, and evolutionary psychology. The traffic accident data will provide a solid foundation in terms of understanding who is killed, in what number, and under what circumstances. This will elucidate how various programming is likely to impact the current traffic accident figures and what is at stake if self-driving cars do not receive widespread acceptance. Data coming out of experimental philosophy provides an insight into people's moral intuitions and preferences, at least in the experimental setting. And evolutionary psychology sheds light on the kind of creatures humans are, in terms of our psychological traits and how we can expect people to react, generally speaking, to issues surrounding self-driving cars. Importantly, it is necessary to bear in mind that our evolved psychology is likely to constrain programming options, with some algorithms being more realistic, in terms of public acceptability, than others.

I have chosen to focus on this problem for an important reason. Namely, our answer is likely to affect whether or not self-driving cars are accepted by the public, which, in turn, will determine whether or not we will reap the benefits that the new technology is set to offer. For instance, if we answer that self-driving cars should *always* protect pedestrians, no matter what, even if this means sacrificing passengers, then perhaps we should not be surprised if the public at large reject buying and travelling in such vehicles. Indeed, it is probable that most parents would be unwilling to purchase a self-driving car which, due to its programming, would sacrifice their child when confronted with a pedestrian, even more so if the pedestrian has, knowing the risks to other road users, chosen to cross the street in an illegal and dangerous manner. It is one thing for a parent to accept a trade-off where the harm that could befall their child is the result of sheer bad luck. Whereas, accepting the potential for harm

stemming from someone being thoughtless, impatient, or even ill-willed, could be too much of a psychological stretch for the vast majority of parents to make. Furthermore, given that algorithms will be developed far removed from the 'heat of the moment' of an accident, by individuals fully conscious of their decisions, we should expect the public to hold those involved in the development and implementation of algorithms to a much higher standard than drivers who simply react. The decision behind developing and implementing a particular algorithm will need to be supported with well founded and compelling reasons, particularly in situations that involve serious injury and loss of life. Otherwise, we may see the prolonged, if not indefinite use of conventional vehicles, which, as outlined above, come with significant costs. A situation that, if we truly are motivated to reduce harm and improve the lives of the public, we should strive to avoid.

A Utilitarian Approach

For this project I am taking a very specific approach. Namely, I will be dealing with the topic from a utilitarian point of view. Although this is not the place for an exhaustive defence of utilitarianism, it is still important to provide a general outline of the theory, discuss why I think it is so well suited for tackling the problem at hand, as well as explore some of its most common criticisms while showing that such attacks not only fail to count against the theory, but actually count in favour of its use when dealing with issues such as the programming of self-driving cars. In doing so, I aim to present a strong case for adopting utilitarianism as the approach for dealing with the ethics of self-driving cars[1].

Utilitarianism is a normative theory in the consequentialist branch of ethics, meaning that is ultimately concerned with consequences. Roughly speaking, the consequences that stem from a particular act (as well as failure to act) or rule are what determines whether something is deemed to be morally good or bad, with a morally good action being one which produces, on balance, more good consequences than bad (Višak 2013: 19). This fundamental principle of the theory is clearly spelt out by one of its founders, Mill, when he states: “actions are right in

1 To make this absolutely clear, and to ensure there is no misunderstanding regarding the scope of this project, here I am adopting a utilitarian approach, and I am treating this topic as a problem within applied ethics. This means that I am assuming utilitarianism, and, as I will go on to say, rule-utilitarianism. Given the constraints of this project, I am not offering a comprehensive defence of utilitarianism or engaging with debates and problems within utilitarianism, nor will I be dealing with other ethical systems in any meaningful way. Rather, I am applying my adopted version of utilitarianism to self-driving car ethics in the Australian context

proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness. By *happiness* is intended pleasure, and the absence of pain; by *unhappiness* pain, and the privation of pleasure [italics added]" (2001: 7). Thus, when evaluating self-driving car algorithms from a utilitarian perspective, our ultimate concern ought to be with the consequences they will produce in terms of impacting people's happiness, where the promotion of happiness is judged to be *good*, and the reverse, *bad*.

In quoting Mill on what it is that determines whether something is morally good or bad, I have referenced happiness, which is one of several properties that a utilitarian could regard as what ought to be valued. This kind of utilitarianism is called hedonistic utilitarianism. When speaking of consequences, and making decisions so as to bring about the best possible consequences, we need to be clear on what it is that makes consequences good or bad, and hedonism holds that pleasure is the only thing that is *intrinsically* good, and that only pain is *intrinsically* bad (Bradley 2009: 4). Thus, utilitarians of the hedonistic variety, such as Mill and Sidgwick, aim to produce states of consciousness "in which there is a surplus of pleasure over pain", and to move from what is good for an individual to what is good for the broader public, the hedonist is therefore interested in the sum total of happiness for all those who can be affected (de Lazari-Radek & Singer 2014: 240).

Given that the ethics of self-driving car programming deals with large groups of people, it is practically impossible to make fine-grained utility comparisons between all such individuals. Moreover, people often have vastly different ideas and beliefs about what makes life worthwhile which are hard to compare. Some may view knowledge as what is of ultimate value, whilst others may contend that the satisfaction of preferences is all that counts. Adopting hedonism can make life much simpler when dealing with large groups of people. Arguably, this is because hedonism subsumes all such concepts of value, and sees them merely as *instrumentally* good. That is, they are good in so far as they are a means to achieving what is *intrinsically* good: pleasure. We may value things such as knowledge or the satisfaction of our preferences, but we do so because of what they result in. For instance, we may get pleasure by coming to gain knowledge that allows us to understand something, not to mention that, on the whole, knowledge tends to make our lives go better by affording us more opportunities, and leaves us better situated to avoid pain inducing mistakes. It is knowledge that has enabled us to develop things such as sanitation, modern medicine, central heating, as well as self-driving cars. All of which improve peoples' lives by enhancing the prospect that

they can live in a manner that makes them happy. Conversely, it would be difficult to assert that knowledge or preference satisfaction is good if it caused pure unhappiness. This is not to say that pain, although intrinsically bad, cannot give rise to good, instrumentally speaking (de Lazari-Radek & Singer 2017: 42). However, if knowledge or preference satisfaction resulted in *nothing* but pain, and produced no pleasure, not now or in the future, then, the hedonist argues, it would no longer be valued.

To further illustrate utilitarianism, let us turn to a scenario where it gives clear direction. In his well known 'drowning child' thought experiment, the utilitarian philosopher Peter Singer (1972) asks us to imagine that en route to work you pass by a small pond. As you walk past the shallow pond you notice a young child flailing about, struggling to keep their head above water. You look around for the child's parent or guardian, but there is no one. That being the case, you quickly realise that if you do not wade into the pond and help the child, they will drown. Although going into the water is perfectly safe for you, the thought crosses your mind that doing so will ruin your new and expensive shoes and suit. The costs associated with ruining your shoes and suit, or being late to work, pale in comparison to the life of a child. As such, you ought to save the child. Failing to do so would be morally wrong.

From a utilitarian perspective, you would have to provide a strong reason for not saving the child, something much more compelling than evoking your fancy new pair of designer shoes. In other words, if something comparable to the loss of the child's life was at stake, then you might be justified in not saving their life. We could suppose that, just after noticing the child in the pond, a car slams into a nearby lamp post turning the vehicle into a crumpled wreck, with a small fire breaking out in the car's engine bay. The driver is unconscious behind the wheel, whilst two children in the back seat cry out. Just as no one is nearby to help with the child in the pond, you are the passengers' only hope. Here you face a choice: save the child in the pond, or, save at least two, possibly three, passengers from the wrecked car before the fire grows too large for you to safely rescue them from being burnt to death. However, much like wading into the pond, saving the car passengers will impose a cost on you by ruining your new suit. In such a scenario, utilitarianism would not only accept abandoning the child in the pond, but would direct you to save the greater number of lives from the burning vehicle. This is not to say that utilitarians think that drowning children are no big deal: the panic and death of the drowning child *are* significant negative consequences. However, all things being equal, saving three lives would be a far better choice, and would clearly result in greater good

consequences compared to saving just one.

Although the basic concept of utilitarianism is quite intuitive and very easy to grasp, there are some key elements that are important to make note of. Using Mill's account, it is clear that one must be able to experience states of consciousness in order to fall within the utilitarian scope of moral concern. Indeed, if something cannot feel pleasure or pain, then it cannot be wronged. Or, if an action fails to produce any consequences for those whose conscious experience can be impacted, then it cannot be judged from a moral point of view. As I have argued elsewhere (Bennett 2018), although we may value an object such as a rock for some reason, given that it lacks the capacity for experience and as such cannot be harmed, kicking it is an act that falls outside the moral realm (Jaworska & Tannenbaum 2013). However, if kicking the rock results in it striking a passer-by, causing them pain, then, all things being equal, from a utilitarian point of view this act is morally bad.

Furthermore, unlike some other moral theories, utilitarianism does not rule-out or forbid any acts themselves. Even though some ethical theories have strict rules regarding things such as killing, lying, or using others as a means to an end (here I have in mind deontological theories such as Kantian ethics), utilitarianism would potentially allow, or even require us, to engage in or promote such courses of action, depending on the predicted consequences that doing so, or not doing so, would produce (Višak 2013: 19). For a utilitarian approach, it would be meaningless to claim that permitting a self-driving car to sacrifice person X is wrong because it is wrong to kill or use people, without being able to show that killing or using person X would lead to more negative consequences than not doing so would. Moreover, arguing for or against a particular algorithm by evoking such things as tradition, social norms, or legal doctrine, for example, is incoherent for the utilitarian, unless the arguments can explicitly draw upon or be cashed out in terms of consequences. To put it another way, it needs to be made clear how something will, in some way, affect conscious experience.

Aside from my belief that utilitarianism in some form is the most justifiable, if not correct, moral system (a claim beyond the scope of this project), there are good reasons for choosing to tackle the ethics of self-driving cars from a utilitarian point of view. As with any issue that stands to impact the public at large, conversation on self-driving cars overwhelmingly turns on consequences. Amongst other things, there is significant interest in their promise to greatly reduce traffic accidents (Crew 2015; Lafrance 2015; Marshall 2017), reduce the burden on

medical resources (Bertalan 2016; Gehrie & Booth 2017), and improve traffic congestion (Brown 2018; Childers 2018; Leong 2018). Opposition to self-driving cars focuses on possible negative consequences, such as decreasing individuals' autonomy once they are no longer free to drive as they please (Burguete 2016; Millar 2014; Moor 2016), and questions surrounding the security of self-driving car systems and whether they could be hacked into by those intent on using them to cause harm (Campbell 2018; Garfinkel 2017; Warncke 2018). Furthermore, studies indicate that when it comes to moral problems involving harm, people are indeed sensitive to the numbers: they seem to endorse action that causes the least amount of harm and promotes the most amount of good (Bonnefon et al. 2015; Bonnefon et al. 2016). Consequently, given that the public, as well as those directly involved in legislating the programming of self-driving cars, are likely to argue for or against proposed details of their deployment in terms of the good or bad consequences that they will produce, it makes sense to investigate the ethics of the new technology using a moral theory that is focused on consequences.

On top of all that, in Australia, it is the government that is ultimately responsible for regulating the safety features of all new vehicles (RSA 2018). For this reason, the programming of self-driving cars will essentially become an issue of public policy, and given the specific circumstances of public life, the features of utilitarianism make it well suited for dealing with the topic. Even more, the typical criticisms that utilitarianism faces when it is seen as a personal philosophy actually become its strengths when it is applied to public policy issues, a view which has been well argued for by Robert Goodin in his book, *Utilitarianism as a Public Philosophy* (1995). For instance, utilitarianism is often criticised for being too *impersonal*, in that it requires us to take, as much as we can, an impartial approach to problems. In the words of the utilitarian Sidgwick (1874), we should adopt 'the point of view of the universe', meaning that we ought to stow the baggage that is our individual idiosyncrasies, tastes, biases, and attachments, when assessing ethical issues (Goodin 1995: 8-9). However, in terms of personal conduct, this is often seen as making unreasonable demands on people, ones that direct them to disregard or limit the importance of the most significant aspects of their lives. Namely, their close relationships and the commitments and obligations that intimate attachments bring.

It should be pointed out that, although this is often a favourite criticism for many opponents of utilitarianism, not all utilitarians see their theory as one which demands that people neglect

or end their close relations and the special treatment that they bestow upon such individuals. Instead, it can be argued that, given these tendencies are such an integral, even hard-wired component of human psychology, life would scarcely be worth living if we tried to undermine them. That being the case, if we are seeking to produce the most good, utilitarianism should make allowances for, and even encourage us to promote and foster close bonds, and would accept us giving preference to individuals with whom we have strong and meaningful ties (Greene 2013: 254-285).

Nevertheless, while those who level this complaint against utilitarianism typically hold that people should be able to maintain and give preference to close relationships and the commitments that stem from them, the same cannot be said when we are talking about individuals who are in some way involved in the development of public policy. In fact, the problem of impersonality that utilitarianism is said to face when we are considering conduct within our personal lives, transforms into one of the theory's most important strengths, and even a requirement, when we turn to examine an issue such as the programming of self-driving cars. Although many would be of the opinion that a person should not be morally condemned if they choose to save the life of the own child over the life of a stranger's child, or because they gave preference to helping an elderly member of one's small village over some unknown elderly member of a neighbouring village, such leniency would not be shown to a developer of public policy who crafted policy, fully or in part, for the purpose of satisfying their own personal interests. Indeed, we expect those involved in public policy to put their personal baggage aside. More specifically, we demand that they do not play favourites, but instead serve all members of the public (Goodin 1995: 8-9).

Recall the problem that lies at the heart of this project: a self-driving car dilemma which requires us to decide, who to protect, or conversely, who to harm, in the case of unavoidable accidents in which all courses of action will lead to somebody being harmed. But instead of thinking of the dilemma as involving a self-driving car and a wayward pedestrian, substitute the pedestrian for a cyclist. Now imagine that someone involved in developing the public policy of self-driving car programming does *not* take 'the point of view of the universe'. Rather, they allow their preferences and other personal baggage to play a role in their decisions. In this imagined situation, the individual in question has a strong dislike of cyclists. He remembers an incident in which a cyclist miscalculated the size of the gap between his car and the kerb, which resulted in his car – his pride and joy – being badly damaged, with the

cyclist speeding away to avoid any consequences for his lax judgement. Looking back on this event arouses thoughts that cyclists are malicious people who lack respect for car drivers, but even more, that they deserve some kind of penalty for all the infractions that they have, and will likely, commit. Unsurprisingly, such a failure to take an impersonal approach to the problem is sure to see self-driving cars programmed to give little regard to cyclists, possibly even all vulnerable road users, without any real concern for how this will impact the public at large: the very thing which public policy makers ought to be most concerned about. Under such circumstances, it is clear that utilitarianism's directive that we take an impersonal approach to issues, counts strongly as a reason why it is well suited for dealing with the topic of self-driving car programming. Moreover, it is something that can easily be agreed upon by both supporters of utilitarianism, and its strongest critics alike.

It is also objected that utilitarianism is *coldly calculating* in terms of consequences, and much like the charge of being impersonal, many see utilitarianism's coldly calculating nature as a fault when it comes to personal affairs. By coldly calculating is meant some type of emotional blunting, where people would suppress their feelings and merely make cold-blooded calculations that direct their decision making (Goodin 1995: 9). That said, it is safe to say that most people want to have relationships with those who display strong emotional bonds which lead them to favour (up to a certain point) those they care about, bonds which produce immediate and somewhat stereotypical emotional responses. An individual who is seen to coldly calculate their options in regards to how they deal with the people they are in relationships with is quite likely to evoke some degree of revulsion. In fact, we would probably think that such a person must be suffering from a type of psychological or emotional deficiency. The last thing we want is for our loved ones to view us as a potential site for producing varying amounts of happiness, while they evaluate the likelihood that they could produce more happiness elsewhere. For instance, it is surely not very high on the list of most peoples' desires to have a partner that, when faced with your upcoming birthday, makes it a habit to calculate whether there are other courses of action by which they could produce more happiness in the world, instead of spending their time and money taking you to a cafe for a celebratory birthday lunch. Indeed, we would be inclined to avoid such a person, while admonishing others to do the same.

Just as they responded to the criticism of impersonality, utilitarians can similarly argue that, although their theory may at first glance appear to endorse such a calculating approach to all

aspects of life, this is not the case. When it comes to our relationships, given what we know about humans' social needs, we have good grounds for recommending that people foster strong emotional ties, and that they let the associated emotions provide a rough and ready guide to their behaviours and choices in this realm. However, when talking about issues of public policy, we want choices that will impact the public at large to be undertaken by thoughtful individuals who rationally consider the facts, not by those whose emotions rule their judgements, nor by those who make choices in spite of the consequences (Goodin 1995: 9). It has been well established that humans come with a moral psychology which features a set of typical automatic emotional responses that are induced when confronted with moral problems. Generally speaking, humans have evolved strong mechanisms against causing harm and killing, and for most people, overriding these automatic 'gut-reactions' takes real effort (Greene 2013). But overriding such emotions is precisely what we expect from policy makers. When dealing with issues involving serious harm and the deaths of members of the public, cold calculation is the approach that allows one to go beyond the emotionally confronting nature of the problem, so as to build a clear idea of exactly what is involved, and what is likely to occur with the implementation of various policies. Or in terms of self-driving cars, a cold, calculating evaluation facilitates a thorough investigation of what the application of different algorithms will produce when they are faced with accident scenarios, and how these will play out over the long-term. The opposite of this is not an option. That is, it would be unacceptable for someone engaged in shaping public policy to merely follow their emotions so as to avoid feeling uncomfortable.

And finally, utilitarianism is *consequentialist*, which, as previously mentioned, means that it is solely concerned with the consequences that acts or rules produce, not the acts or rules themselves. For personal affairs, such a position makes many people uncomfortable, particularly those who see morality as a list of commandments, with different acts grouped into categories where they are either permitted or forbidden (Goodin 1995: 9-10). In other words, many acts, no matter their outcomes, are seen as unthinkable. Regardless of whether it might make sense (even on utilitarian grounds) to discourage people from actions that radically conflict with well established social norms, it would be, as Goodin pointed out, “simply irresponsible of public officials (in any broadly secular society, at least) to adhere mindlessly to moral precepts read off some sacred list” in spite of the consequences (1995: 10). Public policy makers have an obligation to tackle difficult topics head on, which will often mean engaging with problems and making decisions that fall outside the bounds of what

would be acceptable for a regular citizen to engage with. For example, we do not want individual members of the public to start wandering through hospital wards with an eye towards withdrawing treatment from gravely ill patients who are using large amounts of medical resources. No matter whether they argue that by withdrawing treatment from such patients, resources are freed up which could then be used to aid a greater number of patients with more positive prognoses, we would condemn such behaviour and would expect them to face serious punishment. However, those involved in bioethics are employed precisely to work on issues related to this, where it is necessary to consider, for instance, how to deal with the distribution of medical resources, as well as medical practitioners' obligations to patients who may be terminally ill, unconscious with no chance of recovery, or in serious pain. With finite medical resources, we need policy which will direct medical professionals in their use, particularly as most medical facilities face an unending supply of people in need. Clearly, given the nature of these sorts of issues, it is not always possible to act or make policy where no negative outcomes will arise. As such, public policy makers will often have to make choices in the spirit of the public good, even though some degree of harm cannot be avoided.

In regards to the moral dilemma of self-driving car accidents involving unavoidable harm, public policy makers will *have* to make a choice. Whatever they decide, someone will be harmed, though the available options may produce significantly different long-term consequences. Given that they are tasked with promoting the public good, a policy maker who throws their hands up and proclaims “I will not actively choose who will be harmed”, will swiftly find themselves out of a job. Moreover, opting to abstain from actively choosing some particular algorithm does not mean that the problem has suddenly gone away, nor does it free the policy maker from moral responsibility for the outcome that follows their decision. Such an individual is, to some degree, accountable for how self-driving cars will be programmed to act, and is charged with increasing the welfare of society's members. Thus, the consequences of their actions (or omissions) will in part determine whether or not we will see a significant reduction in deaths, injuries, and suffering, as well as the freeing of vast resources which could then be redirected to do good elsewhere. For utilitarians, there is no meaningful distinction between doing and allowing. Harm is harm, whether of not you choose to program the car to sacrifice its passenger or merely allow the passenger to be sacrificed (Greene 2013: 240). The outcome is the same.

Perhaps one of the most key aspects of taking a utilitarian approach is that the judgements

which utilitarianism ultimately produces depend on empirical facts. Although a utilitarian approach to *any* topic always commits us to aim at producing the most good, it is the facts of the issue that determine how utilitarianism will direct us to achieve this end (Goodin 1995: 21). Indeed, we "cannot decide what ought to be, or know how to bring it about, without knowledge of how things are." (de Lazari-Radek & Singer 2014: 14). That being the case, it is imperative that we are well-armed with the relevant data when making normative judgements from a utilitarian point of view, particularly when grappling with a topic such as self-driving car programming that is dealing with such significant real-world consequences. Accordingly, it is vital to analyse, amongst other things, traffic accident statistics, data with respect to people's opinions regarding potential algorithms, how the role of self-sacrifice in moral dilemmas is viewed, as well as attitudes to sacrificing close others such as family and friends. Moreover, evolutionary psychology provides a theoretical framework for understanding people's reactions and moral intuitions to moral dilemmas in the experimental setting, as well as predicting how we can expect people to behave in relation to self-driving cars and the policy that dictates vehicle behaviour.

As self-driving cars will need wide-spread acceptance in order for society to reap the benefits that the new technology promises to deliver, a utilitarian approach must at all times remain cognizant of the fact that if potential algorithms strongly conflict with people's expectations or moral intuitions, then it is feasible that self-driving cars, and the desirable consequences that follow from their widespread adoption, will never come to fruition. Taking a utilitarian approach to the ethics of self-driving cars means investigating what the theory will prescribe after taking *all* relevant information into account. In doing so, the goal is to try and foresee – as best as one can – the probable consequences that are likely to follow from different courses of action, in this case, the implementation of different algorithms.

With all this in mind, it is evident that utilitarianism provides a dispassionate and pragmatic approach to tackling ethical problems, and that some of its most common criticisms may not be as troublesome as they first appear, if even at all, especially when dealing with issues of public policy. This is made clear if we consider mandatory seat-belt policy from a utilitarian point of view. With seat-belts, although it must be acknowledged that some small number of people are harmed by having to use them, we need to consider the overall impact that seat-belt enforcement has. Fastened seat-belts may lead to some passengers becoming trapped in burning cars in the aftermath of a crash, resulting in them suffering horrific injuries and

traumatic deaths. However, in light of the overwhelming number of lives that seat-belts save, even setting aside the fact that they prevent a significant amount of injuries, utilitarianism would agree with the current policy that we should accept the relatively small amount of harm caused by seat-belts, given that it is greatly outweighed by the significant benefits that their required use provides (Goodin 1995: 63).

Rule-Utilitarianism

Utilitarianism comes in two main forms, act-utilitarianism and rule-utilitarianism. Act-utilitarianism directs us to choose actions that will maximise utility on a case by case basis. Whereas, as the name suggests, rule-utilitarianism holds that we should develop and abide by rules that, if followed by the vast majority of people, will bring about the best possible consequences (Goodin 1995: 16, 60). Here, I will adopt rule-utilitarianism, for reasons principally to do with treating the topic as an applied ethics problem in the realm of public policy. Doing so involves applying the utility calculus to the possible rules surrounding the behaviour of self-driving cars in accident scenarios, and determining which rule will maximise utility, which, in this case, means promoting happiness and reducing suffering (de Lazari-Radek & Singer 2017: 88).

Act-utilitarians often lambaste proponents of rule-utilitarianism with 'rule-fetishism', contending that their insistence on implementing and following rules will lead to situations in which more utility could be achieved by acting against the rule (Goodin 1995: 18). They assert that rule-utilitarianism is a failure on utilitarian grounds, and that act-utilitarianism presents itself as the best form of the theory, one which truly functions to produce the maximum amount of utility on a case by case basis. In other words, “following a rule chosen on utilitarian grounds, when more utility could be achieved by doing something else, cannot be the utilitarian thing to do” (Goodin 1995: 17). However, although this argument looks good in theory, it soon fails when we deal in the real-world of applied public policy.

The key reasons for choosing to adopt rule-utilitarianism for self-driving car ethics come from the distinctive circumstances surrounding issues of public policy. Public policy makers have to manage a great deal of uncertainty, and in spite of this, they have no choice but to make decisions. Goodin refers to these two factors as 'the limits of reason' and the 'argument from necessity' (1995: 17, 62). In terms of the limits of reason, making ethical decisions almost

always involves some degree of uncertainty. We can never be completely sure of the results our choices will produce, and even when we are making a judgement within familiar circumstances, there can always be exceptional cases where things turn out vastly different than how we anticipated. When we look at the differences between private individuals and public policy makers, we find that private individuals tend to have much more complete information regarding the details of their situation and the consequences that the different possible choices might have for them and the people around them. While on the other hand, public policy makers operate on generalities, averages, and aggregates, and are unlikely to know how their decisions will impact specific individuals in specific instances. Nevertheless, the argument from necessity refers to the fact that public policy makers are required to make policy decisions (Goodin 1995: 62-63), and generalities are sufficient for them to put self-driving car algorithms through the utility calculator and choose the rule which will produce the most good for the public at large.

Furthermore, unlike private decisions for one's own life, public policy needs to be publicly knowable, and developing clear rules achieves this. As Goodin states, “rules serve to maximize utility in the real world by being easier to communicate, easier to inculcate, easier to remember, and easier to apply” (1995: 17). If we want the public to embrace self-driving car technology, between the two forms of utilitarianism, we must adopt rule-utilitarianism. Just as consumers want to know if the new car they are looking at will deploy its airbags in the event of a crash, they will also want clear answers in regards to what a self-driving car will do in various circumstances. Rule-utilitarianism will do just this, by providing unambiguous guidelines that the public can rely upon and use to make informed choices.

Experimental Philosophy and Human Psychology

In light of what utilitarianism is, and what is needed in order to make a sound utilitarian judgement, let us now turn to experimental philosophy, before highlighting some important findings that are highly relevant for the project at hand. In *Experimental Philosophy*, Joshua Knobe (2007) gives an overview of the field of experimental philosophy and the ongoing disagreement concerning how work coming out of this relatively new endeavour should be understood in relation to analytic philosophy. Knobe asserts that people's intuitions have often played a significant role in philosophical debates within analytic philosophy, with analytic philosophers routinely making use of people's intuitions in order to untangle moral dilemmas

(Knobe 2007: 81-88). Although intuitions are usually given weight by analytic philosophers, it is often the case that their claims about people's intuitions are not based on sound empirical data. Rather, in many instances, philosophers merely state their case then confidently assert that people's intuitions would coincide with their position. Whereas, instead of merely making assumptions about people's intuitions when tackling ethical dilemmas, those using the methods of experimental philosophy strive to subject claims about people's intuitions to systematic experimentation and statistical analysis (Knobe 2007: 81-86). Accordingly, this approach is not only able to investigate the validity of philosopher's claims regarding people's intuitions, but it also leads to a greater understanding of the mechanisms and processes that are at the heart of such intuitions (Knobe 2007: 88).

As the ethical issue of what sort of algorithm self-driving cars should operate under in the event of an unavoidable accident is one that will be of broad general interest, it is necessary for those working on the issue, especially utilitarians, to form a clear understanding of the public's likely reactions to potential algorithms. If the predicted outcomes of a particular algorithm are anticipated to cause public outrage resulting in consumer rejection of self-driving cars, the potential for harm reduction through the spread of self-driving cars will surely be undermined. The methods of experimental philosophy provide a clear way to avoid merely relying upon one's own intuitions when thinking about how others will react to different algorithms, which in turn will only add clarity and strength to one's argument.

Bonnefon, Shariff, and Rahwan (2016) have taken a data driven approach to understanding people's attitudes towards potential algorithms used to operate self-driving cars. For them, the methods of experimental philosophy are a useful tool for developing a clear understanding of how different algorithms are viewed, and for determining the likelihood that self-driving cars operating under such algorithms will be accepted by the public at large. Interestingly, they report that participants overwhelmingly approve of what are referred to as 'utilitarian' algorithms. That is, algorithms that direct cars to cause the least amount of harm, even if that means risking or sacrificing the car's passengers, so long as the number of passengers is less than the number of pedestrians involved in a given accident scenario. However, there proved to be a very important difference between participants ethically approving of such an algorithm and being willing to purchase or travel within a self-driving car that would sacrifice them to avoid harming two or more pedestrians.

In one study, seventy-six percent of participants stated that when confronted with a group of wayward pedestrians, the moral act would be for the self-driving car to sacrifice its passenger in order to avoid harming pedestrians. When asked to rate the most ethical course of action for a self-driving car using a scale of 0 (protect the passenger no matter the cost) to 100 (aim at causing the least amount of injury/death), subjects overwhelmingly indicated a preference for the upper end of the scale. However, when questioned on whether they would be inclined to purchase and travel within such a 'utilitarian' self-driving car, although they had just morally praised said vehicle, they expressed that they would *not* be inclined to buy or be a passenger in one. Indeed, the reported likelihood of buying a self-driving car that would sacrifice passengers to save a greater number of pedestrians was significantly lower than the likelihood of buying a self-protective model (Bonnefon et al. 2016: 1574). Furthermore, the participants made it clear that they would disapprove of legislation that enforced programming vehicles with such an algorithm, which, as Bonnefon et al. point out, would most likely result in the delay – if not total rejection – of the adoption of self-driving cars. Meaning that many lives would continue to be harmed, and even lost, through the prolonged use of our current conventional vehicles (2016: 1575-1576).

The work of Bonnefon et al. is significant due to the fact that it touches upon the application of utilitarianism to the ethics of self-driving cars. It deals with the development of algorithms that will guide self-driving cars in the face of unavoidable accidents in which harm is inevitable, and highlights people's attitudes towards instances in which cars would either sacrifice their passengers or vulnerable road users. On reviewing the experimental data provided by the authors we are able to get a sense of what is practical in regards to achieving the wide-spread uptake of self-driving cars, which is a crucial factor for using a utilitarian approach to the topic. Regardless of whether they morally approved of a particular type of algorithm when it was presented as a hypothetical scenario, it was made clear that people would not purchase or travel within a self-driving car that would sacrifice them as a passenger. This is noteworthy given that it would be the public who, through their consumer choices, would seem to have the power to either facilitate the wide-spread use, or rejection of, the new vehicles. Another key finding was that people would oppose laws that made passenger-sacrificing cars mandatory. So, even if a government attempted to implement such self-driving cars, there is great risk that the public would simply throw their support behind whichever party or politician asserted that they would not force the use of passenger-sacrificing algorithms. These findings go to show just how important experimental philosophy

is for the topic, and this will be built upon later when we look at work that has included some important details, such as bringing family and friends into the mix of people that may be harmed in moral dilemmas. For now, we would seem to have reason to exercise caution against programming cars to give equal preference to the safety of all road user groups. Instead, if we want to obtain the benefits that come with self-driving cars, a utilitarian would presently be inclined to favour the safety of passengers, even though, at first glance, such a view might seem to be contradictory to a utilitarian algorithm.

Bonnefon et al. have used 'utilitarianism' in a very specific manner when talking about algorithms. Namely, in a very narrow sense that lacks foresight, in that it is solely concerned with individual accident scenarios and weighing the costs and benefits for only those immediately involved. Self-driving cars are referred to as operating under utilitarian principles when they cause the least amount of harm within the confines of a given accident. This means that such cars should, for example, always sacrifice their passengers whenever the amount of non-passengers involved in an incident is greater than the number of passengers travelling within the vehicle. In other words, if a car containing one passenger is faced with two or more wayward pedestrians, then it is deemed to be utilitarian so long as it takes evasive action to avoid the pedestrians, even if the only way of doing so is to crash the vehicle into a wall, resulting in certain death for the car's occupant.

Viewing such cars as utilitarian is, however, problematic. Specifically, viewing utilitarianism in this manner neglects important variables such as typical human psychological traits and foreseeable long-term consequences that are likely to follow from the broad implementation of such an algorithm. Although it is highly useful to develop an understanding of people's attitudes to potential algorithms, including 'utilitarian' (in a narrow sense) algorithms, we should not assume that this is what a car operating under utilitarianism would actually look like. Applying utilitarianism to the development of self-driving car algorithms, or any problem for that matter, requires a much broader view of the consequences. Here, I aim to apply utilitarianism in its full sense, as an ethical system that takes *all* morally relevant factors into account, wherein a utilitarian algorithm will be that which promotes the greatest amount of good consequences compared to other alternative algorithms. This approach will not only take the individuals tangled up in an individual accident into account, but will also give concern to how algorithms are likely to impact society at large, and over time. Subsequently, we should be prepared for the possibility that utilitarianism will prescribe a vastly different

algorithm than one that aims at causing the least amount of harm within the bounds of a specific traffic accident. With this is mind, let us now turn to human psychology and the traits that, generally speaking, humans possess, and how these may influence the programming of self-driving cars.

We humans are evolved animals, and given what we know about our evolved psychology, particularly *parental investment* (Trivers 1971) and *kin selection* (Hamilton 1964), we should expect people, especially parents, to have a strong reaction against a new device that would explicitly put members of their family at risk. However, instead of just stating that people are likely to reject buying and travelling in self-driving cars that are programmed to sacrifice them and their loved ones in certain instances, and relying on findings coming out of the experimental setting as support, it is important to go beyond making superficial claims. Indeed, as this is a practical issue involving real-world harm, claims about what people are likely to reject or accept ought to be supported as best as they can, and providing an account of the evolutionary mechanisms behind such actions will go some way to strengthening the findings coming out of experimental philosophy, as well as the ultimate argument of this project.

Before going any further, it should be made clear that discussing our evolved psychology should *not* be taken to mean that I am in any way drawing moral conclusions from evolution. Evolution is blind, and cares nothing for ethics. Just because something has evolved a certain way does not mean that it is morally good, and the typical evolved moral intuitions held by most people do not necessarily align with justifiable moral judgements. Nonetheless, as was previously stated, in order to make moral judgements we do need a decent understanding of what and who we are dealing with, and this includes developing an understanding of our fundamental nature. Humans are not infinitely free, psychologically speaking. Our minds are constrained, and given our evolutionary history, we should expect people to have tendencies that may make the implementation of some algorithms practically impossible. Thus, in order to foresee how people are likely to react to self-driving car programming, and the consequences such reactions could lead to, we cannot overlook evolutionary psychology and what it tells us about the type of creatures humans are.

The forces of Darwinian natural selection (Darwin 1859) have shaped who we are, not just physically, but also psychologically. According to Dawkins (2006), we ought to view

organisms, including ourselves, as 'survival machines': machines that have been built by genes that have the 'goal' of passing copies of themselves to later generations. And as we share copies of genes with our offspring and kin, we can expect the existence of traits that drive us to protect our genetic relatives (Dawkins 2006: 88; Trivers 1985: 109). This should not be taken to mean that genes are consciously choosing strategies in the same way that one might develop strategies for how they should play to win a game of rugby. Instead, here we are talking about the typical human emotional and behavioural traits that stem from our genotype, which act as mechanisms to promote the continued existence of the genes that underpin them (Buss 2016: 195-196).

As LeVine (1988) pointed out, no matter the culture, it is universal that parents seek to secure the health and survival of their children. Although this is not a moral fact in itself, it is nevertheless a feature of human psychology that we cannot ignore when trying to assess the possible consequences of a decision regarding self-driving car programming, or any public policy issue for that matter. Genetic relatedness tends to predict how willing one is to ensure the survival of others, even at a significant cost to oneself (Trivers 1985: 45). With an increase in genetic relatedness, we tend to see a greater degree of emotional closeness, something that is made abundantly clear in the high level of grief typically experienced by parents at the loss of a child (Littlefield & Rushton 1986). It is these types of emotions that prompt the protective behaviour commonly seen between family members, and in how parents behave in regards to their children.

Reproducing and successfully raising offspring to maturity is a costly exercise, especially for females. Amongst other things, it involves a long gestation, bearing, feeding, as well as nurturing and protecting the child in a bid to ensure that they themselves reach reproductive age (Buss 2016: 104). Parental investment is defined as “any investment by the parent in an individual offspring that increases the offspring's chance of surviving (and hence reproductive success) at the cost of the parent's ability to invest in other offspring” (Dawkins 2006: 124). In this context, survival is critical because parents share roughly half their genes with their offspring. Thus, the survival of one's offspring to sexual maturity (so long as they are reproductively successful) results in the replication of a number of the parent's genes. So, given that offspring are highly important as genetic vehicles for parents, natural selection has selected for parental mechanisms that promote offspring's survival and reproductive success, and this is exactly where strong parent-child emotional bonds come into play. From an

evolutionary perspective, the strong emotional connections between parent and child, including deep parental love, is an evolved mechanism that works to ensure the transportation of one's genes into future generations (Buss 2016: 195-196). Additionally, such behavioural tendencies are typically passed from generation to generation, since those who possess the offspring caring traits are more inclined to successfully raise offspring that will be endowed with the same traits, in turn leaving them well positioned to do likewise.

The same applies to relationships between kin. Just as natural selection favours traits in parents that work to ensure the survival of offspring, our kin share copies of our genes, and behaviour that increases the chances of our kin surviving means that our shared genes are also likely to be passed to later generations. As such, kin selection explains the general tendency for within-family altruism, and the closer the genetic relationship, the more we should expect relatives to act altruistically towards members of their family (Dawkins 2006: 94, Trivers 1985: 109).

It is through our emotions such as love, disgust, and shame, that evolution often 'directs' our behaviour (Greene 2013). In evolutionary terms, the emotions felt by parents are the proximate explanation, or the mechanism, while the survival of our shared genes in our offspring is the ultimate explanation, which accounts for the function of the mechanism (Tinbergen 1963). Importantly, an evolutionary explanation does not make the love felt by the parents any less real or diminish the bond between parent and child, or other family members. In fact, evolution would favour those who are more caring: the offspring of parents who care about them are more likely to survive than those that are ignored or harmed by their parent's choices.

From an evolutionary psychological perspective, if self-driving cars were programmed to give preference to vulnerable road users, we should expect people, especially parents, to be inclined to reject buying or using such vehicles, and having an understanding of evolutionary psychology provides an explanation for the findings of Bonnefon et al. (2016), that people, although willing to morally praise self-driving cars that sacrificed passengers to save a greater number of pedestrians, indicated that they would be unlikely to buy or travel in such vehicles. Moreover, if a mother and child were to be killed as passengers in a self-driving car because it avoided some careless pedestrians, the emotional salience of the event is certain to see it feature prominently in the headlines and promote a widespread backlash against such

programming. This would be so even if the programming resulted in, for the accident scenario in question, the best outcome in terms of number of lives saved. Thus, as it should by now be clear, it is not solely how many lives that will be saved or lost in each individual accident scenario which is what is of utmost importance, but the sustained, long-term consequences that will come about through the implementation of an algorithm.

In *The Role of Self-Sacrifice in Moral Dilemmas*, researchers (Sachdeva et al. 2015) examined people's attitudes towards the various forms of sacrifice in hypothetical moral dilemmas. They point out that although self-sacrificial acts feature prominently across cultures in heroic stories related to national identity, religious teachings, mythology, as well as being used to teach children the notion of virtue through story-telling, self-sacrifice has largely been absent from work on moral dilemmas (Sachdeva et al. 2015: 2). In light of this, the researchers conducted experiments using a modified version of the 'trolley problem', which investigated the extent to which participants endorse self-sacrificial behaviour in hypothetical moral dilemmas, how people respond to options of sacrificing family and friends instead of strangers, as well as how attitudes to self-sacrificial acts change when a subject's view of the moral dilemma shifts from a first-person perspective to a third-person perspective. However, before we move to discuss the findings of such research, a brief run down of the trolley problem is in order.

The trolley problem (Foot 1967; Thomson 1985) is a thought experiment in ethics in which participants are faced with a choice between two options, both of which lead to unavoidable harms. It involves a runaway train hurtling down a track. The train's brakes have failed and there is no way of stopping it. Further down the line five workers who are busy undertaking track repairs stand in the path of the oncoming train, with no possibility of getting to safety: they face certain death. You happen, however, to be standing next to a lever which is used for switching tracks and diverting trains. If you throw the lever the train will be directed down a side-track, saving the lives of all five workers. Still, doing so is not without issue: the side-track is not clear of people. There, a single worker is doing track repairs, and although diverting the train will save the others on the main line, doing so will kill this lone worker. As such, the thought experiment challenges us to think about what action would be the most ethical. Namely, should you do nothing and allow five people to be killed, or do you throw the lever to save the five by killing one? Put more simply, it is five lives versus one life.

Although many have criticised trolley-type thought experiments as being unrealistic and unlike anything we would ever actually encounter (Baumen et al. 2014; Gold et al. 2014), self-driving cars challenge this view. It is clear that the problem at the heart of this project involves deciding between two options that involve harm. The main difference between it and the trolley problem being that those involved in the development of self-driving car policy have the good fortune of not finding themselves confronted with an imminent accident which requires them to make a decision on the spot. Instead, policy makers, after evaluating data and critically reflecting upon the various options, will be able to spend time making rational decisions. Nonetheless, the trolley problem now seems to be real. Thus, experimental work that has been conducted using trolley problem thought experiments offers useful insight into peoples' moral psychology, which in turn provides highly relevant data for the problem at hand.

To return to the experiments, it was found that when presented with the researchers' modified version of the trolley problem, in which there was an option for self-sacrifice so as to avoid harming others, participants preferred self-sacrifice rather than harming a stranger. This is in keeping with the view that humans have evolved a moral psychology that, generally speaking, sees them inclined to avoid harming others (Greene 2008: 43). However, when the moral dilemma also involved close others such as family and friends, it was found that that subjects viewed the sacrifice of a close other as the most morally reprehensible of all options (Sachdeva et al. 2015: 6). Although these findings suggest that self-sacrifice is always a morally praiseworthy action, it was discovered that the participants only preferred the self-sacrificial course of action when viewing the dilemma from a first-person perspective. That is, a shift in perspective to being a third-person observer no longer saw participants express a preference for the 'self-sacrificial' act over a stranger being sacrificed. In other words, participants did not think that someone choosing to sacrifice themselves in order to spare strangers was the preferred course of action (Sachdeva et al. 2015: 8). The authors do, however, acknowledge that there are questions around ecological validity that stem from their work. Namely, we are justified in questioning whether the findings within the experimental setting would carry over into the real world. Although participants indicated a preference for self-sacrifice when presented with hypothetical moral dilemmas, it seems unlikely that people would behave this way in the real-world, particularly in scenarios involving strangers. Indeed, if people really were so self-sacrificial, it is not unreasonable to think that we would see more individuals making far less significant sacrifices in their daily lives, such as giving up factory

farmed animal products, as well as displaying a greater interest in charitable giving.

Findings from previous studies also tend to support that people are unlikely to be so sacrificial, in that they would be unwilling to buy or travel within passenger-sacrificing vehicles (Bonnefon et al. 2016). Moreover, there may be strong context effects, which is well demonstrated in the following scenario: “jumping on a grenade to save five fellow soldiers can easily be seen as a praiseworthy act, yet a healthy soldier donating all his body organs to save the other soldiers will probably be seen not as a hero but as an aberration” (Sachdeva et al. 2015: 10). If we apply this concept to traffic accidents, we can reflect on whether a person would be praised for accepting self-sacrifice for all incidences involving vulnerable road users. It would be one thing for a self-driving car user to consent to sacrificing themselves to spare a child who managed to wander out onto the street, whereas choosing to be killed because four drunken adults have decided to play chicken with self-driving cars is a vastly different story, and may elicit a similar response to the healthy soldier who chooses death so as to become the source of organs for others.

Such findings are pertinent to questions surrounding whether self-driving cars should sacrifice their passengers to save others. Given that a significant amount of people use their vehicles to transport family, including their children, the recognition that people view harming a close other in order to spare harming a stranger as morally reprehensible is particularly important. Research that delves into attitudes towards self-driving car algorithms could be improved, and perhaps be made more ecologically valid, if participants are presented with scenarios that involve travelling in self-driving cars with friends and family members. Moreover, much like the context around the 'self-sacrificial soldier' (Sachdeva et al. 2015: 10), we ought to give context to the pedestrians involved in self-driving car dilemmas. Indeed, subjects in past studies (Bonnefon et al. 2015; Bonnefon et al. 2016; Wächter et al. 2017) may have provided vastly different responses if told that the wayward pedestrians were, in fact, people blatantly disregarding designated pedestrian crossings, embarking across the street while watching videos on their smart-phones, or under the influence of drugs or alcohol.

A 'narrow' view of utilitarianism applied to self-driving cars also ignores the fact that accident scenarios do not exist in isolation. We must acknowledge that road users in all forms have ongoing interactions, and once the rules of self-driving car programming are widely known, people, such as pedestrians, are likely to respond to them with behaviour modification. At

present, pedestrians have a number of disincentives against crossing streets in dangerous and illegal ways. In addition to knowing that there is always the chance of a police officer being within sight, they also know that not all drivers are created equal. Some drivers pay scant attention, others drive at high speeds which leaves them unable to slow down or manoeuvre appropriately, many drive under the influence of drugs and alcohol, and some small portion of drivers are surely just psychopathic. For the impatient pedestrian, reaping the benefit of getting to their destination slightly earlier by darting out across the street in front of oncoming traffic starts to lose its appeal when they consider just who they may be risking running out in front of. However, the so-called utilitarian cars of the recent literature would remove such disincentives, leaving room, if not providing incentives, for pedestrians to behave in ways that would greatly increase the chances that self-driving car passengers will come to harm (Millard-Ball 2016). If vulnerable road users such as pedestrians can be certain that their safety is prioritised, we would see the removal of a major inhibitor against risky behaviour on their behalf, with the burden of risk being shifted to car passengers. Given this shift, those who choose to adopt a new technology that, not only provides a safer environment for other road, but also stands to have significant positive impacts on society, are liable to feel that the potential costs to them and their families is too great. For this reason, this type of programming boosts the prospect that self-driving cars will be abandoned, which, given what is at stake, is the opposite of what utilitarianism would have as its goal.

Consequences Beyond Crashes

The wide-spread adoption of self-driving cars will have significant impact beyond harming or sparing those directly involved in traffic accidents. Although we must look at the ways in which different approaches to programming self-driving cars will affect the number of people harmed and killed in traffic accidents, it is critical that we look at the consequences beyond the immediate ones of a given accident scenario. Such external consequences are likely to carry considerable weight, and may provide all the more reason to ensure that the public's acceptance of the new vehicles is obtained.

In terms of consequences beyond traffic accidents, Gehrie and Booth (2017) draw attention to the potential far-reaching implications that the adoption of self-driving cars may have in the medical field. Specifically, they bring to the fore the positive consequences that self-driving cars could have on blood banks and transfusion medicine. In their own country, the United

States, it is estimated that the improved traffic safety that is anticipated to result from self-driving cars will see approximately 32,000 less deaths per year. With the lower trauma rates that are expected to follow the uptake of self-driving cars, some within the field of transfusion medicine see their implementation leading to a significant reduction in the demand for blood transfusions, thereby lowering the pressure on the blood supplies held by ambulance teams, emergency departments, and operating rooms. Thus improving the availability of blood to both medical and surgical patients, especially during holiday periods. Moreover, they foresee an additional benefit of having greater financial resources at their disposal for trauma related clinical trials, not to mention a general reduction in pressure on medical resources, broadly speaking (Poczter & Jankovic 2014: 9-10).

Moving beyond the medical field, it has been identified that self-driving cars stand to have a major impact on land usage, given that they will allow for a significant redesigning of parking facilities (Nourinejad et al. 2017). According to Mitchell (2015), a typical vehicle spends around ninety-five percent of its life sitting in a parking spot, as such, parking infrastructure is a key part of urban and transportation planning. Currently, human driven vehicles require parking facilities that have spacious lanes that give drivers room for error, parking bays need to be big enough for drivers and passengers to be able to open doors in order to get in and out of vehicles, car bays must be laid out in two rows so that vehicles are not blocked in, and in terms of large multi-level parking garages that one finds in city centres, apartment complexes, and commercial buildings, such facilities require enough space to house elevators and stairwells. To give an idea of just how much land is used for the purposes of parking vehicles, in the United States almost 17,000 square kilometres of land is devoted to parking, which is an area larger than the Greater Sydney region: 12,367 square kilometres (Nourinejad et al. 2017: 110).

Researchers at the University of Toronto (Nourinejad et al. 2017) have investigated the optimal design of self-driving car parking facilities and estimated that, compared to current designs, the average space needed per vehicle within parking facilities could be reduced by two square metres. This can be achieved by, for instance, creating passenger drop off zones at the entrance to parking facilities where people can exit and enter vehicles, which would eliminate the need for space within the facility for people to move around in, including stairwells and elevators. From there, self-driving cars can then be directed by either car park operators or with automated systems to their parking space, and as self-driving cars will not

make driving errors, lane size can be kept to an absolute minimum. And finally, because passengers will not be exiting or entering vehicles while they are in parking bays, the space that is necessary for opening doors is no longer needed, which will allow for cars to be parked extremely close together. Consequently, Nourinejad and colleagues posit that such facilities could reduce the amount of car park space needed by around sixty percent, and at best, by eighty-seven percent. Importantly, if space that was previously needed for parking conventional vehicles becomes freed up, it can then be used in socially beneficial ways. That is, it can be used for residential and commercial purposes, or it could even provide land for public spaces such as parks or green areas within inner cities (Nourinejad et al. 2017: 110).

The widespread adoption of self-driving cars also has great potential to reduce energy consumption and emission production (both greenhouse gases and pollutant emissions), which, when we look at the issue of climate change, becomes an extremely important factor for ensuring that self-driving cars get broad public acceptance. Although energy usage and emission production have local effects, their global impact should not be underestimated. Every time we drive fossil fuels are burnt, thereby releasing carbon dioxide and other pollutants into the atmosphere which has an impact on the world's climate (Singer 2011: 216). This may be directly, through the operation of internal combustion engines found in our current conventional and hybrid vehicles, or through the generation of electricity needed to charge electric cars, which, in Australia, overwhelmingly comes from the burning of fossil fuels (DOEAE 2018).

We now know that the atmosphere has a limited ability to absorb gases without there being negative consequences, and we may in fact already be pushing our atmosphere beyond its capabilities. At the end of the industrial revolution, the amount of carbon dioxide in the atmosphere reached a level of 390 parts per million, and this level is currently rising by roughly two parts per million every year. The general agreement amongst scientists has been that, if our average temperature increases by 2 degrees Celsius, then we will face significantly dangerous consequences. Such an increase is thought to come about if we reach something like a carbon dioxide level of 450 parts per million, a figure that, based on current trends, will be reached by 2040 (Singer 2011: 218-219). In fact, some, such as members of the U.S. National Aeronautics and Space Administration, have asserted that, in order to maintain a planet conducive to healthy life, we must reduce carbon dioxide levels to 350 parts per million, a level that we have already surpassed (Hansen et al. 2008). Importantly, the World

Health Organization (2004) has claimed that the planetary warming which we have already experienced since 1990, resulted in 140,000 additional deaths in 2004. And in 2007, a scientific group set up by the United Nations Environment Program and the World Meteorological Association showed that if the average global temperature were to rise by two or more degrees, we would see the world's water resources under considerable pressure, not to mention that such an average temperature increase would expose at least sixteen million people to the horrors of coastal flooding, thus triggering large scale humanitarian crises (Singer 2011: 217).

Now that the significance of climate change has been made clear, let us turn to work that has looked at how self-driving cars may differ from conventional vehicles in terms of energy use and emission production. Using a method that, roughly speaking, takes 'snapshots' of vehicle operation under different circumstances, researchers (Barth & Boriboonsomsin 2008) have been able to create an energy/emissions model that maps energy and emission values as a function of average traffic speed (Barth et al. 2014: 104-105). In doing so, it has been found that energy use and emission production is high at low average speeds, with the reason being that at lower speed it takes longer to reach an intended destination, meaning that vehicles are on the road for a longer period of time. While at a more midrange speeds of around seventy kilometres per hour, there is a tendency for energy use and emission production to slow down, with energy and emissions increasing once vehicles are travelling at around ninety kilometres per hour and above. The increase in energy use and emissions output at higher speeds is the result of aerodynamic drag, as higher drag places greater pressure on a vehicle's engine leaving it needing more energy to maintain speed, leading to the emission of more carbon and other pollutants (Barth et al. 2014: 105). From these general findings, Barth, Boriboonsomsin and Wu (2014) have identified three areas where self-driving cars are able to better conventional vehicles in terms of reducing energy use and emission production. Namely, by reducing traffic congestion, smoothing traffic flow, and introducing platooning. Each of which I shall now explain.

With human driven vehicles, traffic conditions often deteriorate because of human behaviour. For example, it is common that an act such as lane merging suffers from people being indecisive, having delayed reaction times, incorrectly matching speeds, or being too aggressive. Because of this, we routinely find traffic coming to a crawl, if not a standstill, leaving roads congested, increasing energy use and emission production. Whereas, given that

they are anticipated to outperform human drivers, self-driving cars will be able to manoeuvre much more accurately whilst maintaining speeds. Furthermore, with better reaction times, strict adherence to traffic rules and speed limits, and inter-vehicle communication, self-driving cars will be able to improve traffic conditions by reducing the stop-start traffic movements that often results from poor driving, as when people enter onto roadways in a manner that forces others to brake sharply, or when drivers fail to reach the speed limit and maintain a steady speed. Consequently, self-driving cars should leave traffic flowing much more smoothly, with vehicles maintaining average, and even higher, speeds (Barth et al. 2014: 108-110). And they will allow for a higher capacity of cars on the road with less congestion, with a reduction in energy requirements and emissions output (Barth et al. 2014: 106-107).

Additionally, the possibility for self-driving cars to be able to safely follow closely together at speeds leads to the potential for platooning. Here, think of the phenomenon of 'slipstreaming' in motor racing or cycling, where one or more cars or bikes follows very closely behind a leader. The leader leaves a wake of air behind that has a reduced pressure. Those behind are then able to travel through this area of lower pressure air, which requires less energy due to less aerodynamic drag. Interestingly, it is not only the followers that reap the benefits of platooning, it has been found that lead vehicles also benefit, with a reduction in energy use of around ten to fifteen percent (Browand et al. 2004). On less congested, more smoothly flowing roads, with traffic maintaining average speeds, travel times should fall, which will not only decrease energy use and emission production, but will also decrease costs associated with running a vehicle. This could also improve people's well-being, as it has been found that longer commute times tend to negatively impact how satisfied people are with their lives (Hilbrecht et al. 2014).

In outlining ways in which the advent of self-driving cars are likely to affect fields related to transfusion medicine, land and energy use, as well as the global consequences that are associated with climate change, it is clear that any investigation into the ethics of self-driving car programming needs to look beyond the scope of individual traffic accidents and the people directly involved in them. We must give consideration to how different algorithms can have far-reaching consequences. This is especially the case in terms of a utilitarian approach to the topic, for which the consequences are what ultimately matter. Here, I have touched upon a few factors external to traffic accident scenarios that provide extremely weighty reasons why striving to bring about the acceptance of self-driving cars is so important.

Accordingly, this should encourage us to evaluate where else self-driving cars may lead to other significant positive or negative consequences, and it reinforces that an ethical approach which focuses on consequences should be adopted for tackling this issue. To neglect such serious consequences, would, particularly for the public policy maker, be abhorrent (Goodin 1995: 4). In fact, given what is at stake, and in light of the impact that self-driving cars could have on some serious issues, a non-consequentialist approach to self-driving cars may actually be unethical.

Taking a Broader View

When we are grappling with moral questions, particularly those that involve the suffering of others, it is unsurprising that one may feel a sense of urgency pushing them to act or make decisions with all possible haste. However, we should not be too quick to make utilitarian normative judgements until we have thoroughly considered whether our prescriptions will actually produce the best consequences. Indeed, we want to be confident that we are not actually going to make matters worse. For example, an aid organisation might make us aware of a far-away village whose population is suffering given the food shortages they are experiencing. The aid organisation's brochure shows a picture of a hungry child with teary eyes and a distended belly. We are left feeling that donating to the organisation, whose sole mission is to distribute bags of rice throughout the region, is the best way to reduce suffering, and conversely, promote happiness. However, this imaginary aid organisation lacks any sort of long-term view of the problem. The extent of their plan is to distribute the bags of rice, with no serious thought as to how their approach will impact the people of the village over time. This is not to say that giving food to hungry people is a bad thing, but if we, as utilitarians, are looking to promote the most good, then we should not necessarily proceed with, or support, the most immediate and emotionally compelling proposal that strikes us. Instead, we ought to weigh the costs and benefits of the different possible courses of action, including over the long-term, and strive to get the most 'bang for our buck', so to speak. It may be the case that giving the villagers bags of rice will result in them becoming content with and dependent on the aid distribution, doing little to improve the villagers' future prospects beyond avoiding starvation. Whereas, after some research, we may discover an organisation whose mission is to provide, along with a smaller quantity of rice, education and farming materials to the villagers, which will enable them become self-sufficient over time, thereby improving their chances of avoiding food shortages and the suffering that follows. In

turn, this plan will enhance the future prospects of the villagers, thus producing a greater amount of good overall.

Although the latter of my imagined aid organisations will produce more overall good than the former, it is true that their plan does involve a trade-off. That is, unlike the action of the former which will provide sufficient rice to satiate all members of the village, immediately reducing suffering to a great degree, the latter will provide less rice, which, consequently, will do far less to reduce present suffering. This is due to the fact that a portion of their resources are used for education and farming supplies, in a bid to address the underlying cause of the villagers' problems. Under such a scheme the villagers will have to devise a distribution system whereby some members of the group, say, the farm workers, may receive a greater share of the limited rice, leaving some, if not the entire population, still facing varying degrees of hunger in the short-term. However, when we take a much broader view, which I believe is the view utilitarianism requires us to take, we can see that given the different results both aid organisations are likely to produce, it is worth making the trade-off where people will initially continue to suffer to some extent, in order to promote a much greater amount of good, even though it will be postponed. As with this imagined scenario, it is highly unlikely that there will ever be a perfect solution (or algorithm) for self-driving cars, though this should not stop us from moving forward with the problem. In the real-world, things happen quickly, variables change, and we can only do our best to predict outcomes in an imperfect way. As Hardin (1988: 17) asserted, “any argument that turns on perfect information, perfect calculation, and perfect theory is a house of cards, [and therefore] is almost entirely beside the point for a practical morality”.

Whilst a broad view of the consequences in the scenario described above uncovered no reason to think that accepting some amount of suffering in the short-term would not be outweighed, this is not always the case. Given enough time, or depending on the details, an act or policy behind an initially promising trade-off may result in significantly bad outcomes. Furthermore, we need to take great care in assessing how people will respond to changes in circumstances. To highlight this, which I consider to be crucial in arguing for a utilitarian self-driving car algorithm, let us turn to Harris' (1975) *The Survival Lottery* and Singer's (1977) well thought-out reply. Harris outlines a hypothetical program which he refers to as the *survival lottery*. The program, roughly speaking, consists of individuals consenting to be part of a scheme which provides healthy organs for those who are facing death due to some type of organ failure or

untreatable disease. Whenever several people are in need of organs, a lottery is drawn, and the individual whose name is called is, with their consent, sacrificed, in order that they become an organ donor. The harvested organs are then transplanted into the bodies of several other individuals who would otherwise die without receiving the healthy organs. For Harris, the point of the survival lottery is that it will save more lives than are lost via sacrifice, given that one person can supply multiple organs which will profoundly benefit a number of people. Furthermore, we will see an increase in the average life expectancy of the participating population, which, Harris argues, all amounts to an upsurge in happiness, making the trade-offs involved worthwhile. That being the case, the survival lottery is a logical application of utilitarianism, which followers of the moral theory ought to endorse (Harris 1975: 86).

The utilitarian philosopher Peter Singer responds to Harris, and, although he agrees that at first glance a survival lottery appears to be "utilitarian planning carried to a new extreme" (Singer 1977: 219), he asserts that it faces an insurmountable problem, one that is ultimately of a utilitarian kind. Namely, it shifts the consequences of imprudent action from the imprudent individual to the rest of society. Under Harris' program, the individual who is irresponsible and careless in their health choices can enjoy the benefits of satisfying their unending desire for cheeseburgers and cigarettes, amongst other things, yet not have to worry about the consequences of diseased organs and a shortened life. Furthermore, the glutton's organs may end up so unhealthy or even diseased, that they become unfit for transplantation, leaving such individuals even less likely to be called for self-sacrifice. On the other hand, the healthy and restrained individual will continue to run the chance of being called for self-sacrifice in order to provide healthy organs for others, even though it is highly improbable that they will themselves require new organs (Singer 1977: 219). Hence, we see a deterioration of the incentives to be healthy and thoughtful. In fact, Singer points out that the scheme will leave healthy individuals bearing the burden of providing organs when a group of gluttons have pushed their bodies too far (Singer 1977: 219). Under such circumstances, individuals who are restrained in their eating habits and dedicated to regular exercise will have little reason to be involved in such a program. And if the survival lottery were to become public policy, prudent individuals would have an incentive to duplicate the care-free lifestyle of the glutton, in order to reduce the probability of being called upon to die. Consequently, such a situation is liable to lead to a total degeneration of health, resulting in more disease, deaths, and a shorter life expectancy than before the lottery began (Singer 1977: 219).

Singer's response is highly important for developing a utilitarian approach to the ethics of self-driving cars. In fact, in light of Singer's paper, the 'utilitarian' self-driving cars of some recent papers would seem to promote a severe form of risk pooling, in which vulnerable road users, such as pedestrians, would lose a significant disincentive against behaving imprudently when crossing roads. That is, if self-driving cars are programmed to prioritise vulnerable road user safety, some vulnerable road users, much like the gluttons, will see the chance of arriving at their destinations quicker by disregarding traffic rules, while bearing little risk of being injured by a vehicle. Consequently, passengers of self-driving cars will face the possibility of being sacrificed for someone else's impatience and thoughtlessness. Not to mention that such programming opens up the possibility for those with bad intentions to exploit self-driving cars for the purpose of deliberately causing harm, a factor which, given recent world events in which terrorists have used vehicles to commit atrocities, should not be overlooked (Bigelow 2016; FBI 2014; Lewis 2015). Because of this, there is great potential for those considering buying self-driving cars to reject such technology and opt to stick with conventional vehicles. Given that research (Bonnefon et al. 2016) has shown people favour self-protective algorithms for themselves, and that they see the sacrifice of a family member or friend as reprehensible (Sachdeva et al. 2015), it is too great a burden to ask people to accept the cost of injury and death for them and their loved ones, when that would include scenarios involving careless or impatient vulnerable road users, as well as religious fanatics hell-bent on causing destruction. It is plausible that such 'utilitarian' algorithms would actually produce consequences antithetical to what utilitarianism would strive to realise through the wide-spread acceptance of self-driving cars.

Traffic Accident Data: Numbers and Circumstances

Maintaining a focus on Australia, let us now turn to the details of traffic accidents. Doing so is crucial for this project, as it allows us to develop a clear understanding of who is injured and killed, in what numbers, and under what circumstances. This, in many cases, will include being able to establish who is responsible for traffic accident casualties. Subsequently, we will then be positioned to compare fatality and injury rates across road user groups and consider how programming self-driving cars to give preference to different road users within accident scenarios could affect these figures. Moreover, it is important to gain insight into the circumstances of traffic accidents, as, when we view the data through the lens of evolutionary psychology, we find that not all traffic accidents are created equal. By this I mean that, when

the causes of accidents become known, people's level of sympathy for those harmed and the degree of risk they are willing to accept to reduce such harm, is likely to change significantly, which, if we want to avoid public rejection of self-driving cars, is something that cannot be ignored when it comes time to decide how self-driving cars should be programmed.

In 2016 (BITRE 2017) there were 1,295 people killed in road traffic accidents in Australia. Of these, 833 were vehicle occupants (either drivers or passengers), and 459 were vulnerable road users (pedestrians, motorcyclists, and cyclists). When we break down vulnerable road users into their different groups, we find that 182 were pedestrians, 248 motorcyclists, and twenty-nine cyclists. In terms of injuries requiring hospitalisation, of the 35,552 people hospitalised, 16,702 were vehicle occupants, while the number of vulnerable road users hospitalised was 17,539, of which 2,562 were pedestrians, 8,335 were motorcyclists, and 6,642 cyclists.[2] Although there was an increase in fatalities of around seven percent compared to 2015, the annual death toll has remained fairly stable in recent years, with a general downward trend over the last decade, excepting the previous three years. Whereas, in regards to serious injury, there has been a gradual increase in the number of road users hospitalised since 2002 (BITRE 2017: iii-iv). In view of such data, it is clear that more vehicle occupants are killed in traffic accidents compared to vulnerable road users. However, vulnerable road users outnumber car occupants when it comes to being hospitalised.

That said, in relation to both fatalities and hospitalisations, the number of motorcyclists involved stands out. Specifically, motorcyclists account for approximately fifty-four percent of vulnerable road users killed, and around forty-seven percent of hospitalisations, which is more than six times the number of pedestrians injured. One startling detail regarding motorcyclist deaths is that roughly twenty percent of all the motorcyclists killed in traffic accidents did not hold a valid motorcycle license. Moreover, around ten percent died whilst not wearing a helmet, and of those who were wearing a helmet, approximately twenty percent were wearing incorrectly fitted helmets (Johnstone et al. 2008: vii). Furthermore, it has been found that about forty-two percent of fatal motorcycle accidents are single vehicle crashes, meaning that no other vehicle was involved in the accident, with the majority of these being scenarios in which riders had simply lost control on a bend. Between 1993 and 2003, the main factor attributed to such fatal accidents was excessive speed, which accounted for seventy

2 It may be noticed that these figures do not add up to 100%. This is due to the fact that some vulnerable road users, such as horse riders, are not included, and the road user type for some traffic accident casualties is unknown.

percent of all single vehicle motorcycle crashes, with the next biggest factor being motorcyclists riding under the influence of drugs and alcohol (Johnstone et al. 2008: 10-19).

In multiple vehicle crashes in which motorcyclists were killed, it has been determined that in the majority of cases the motorcyclist was to blame, with responsibility allocated to riders fifty-five percent of the time. The circumstances surrounding these crashes were similar to single vehicle crashes, with the main factor allocated to motorcyclists being excessive speed, which accounted for around forty-one percent of such accidents, with drugs and alcohol being at play in twenty-one percent of crashes. In twenty-nine percent of cases, blame was attributed to the other vehicle, while both motorcyclists and other vehicles were deemed equally at fault thirteen percent of the time. When another vehicle was to blame for a motorcyclist's death, human error on the part of the driver was the reason cited, with nineteen percent of crashes with a known cause occurring because the driver did not see the rider, while nine percent of such accidents transpired because the driver failed to give way to a motorcyclist (Johnstone et al. 2008: 19).

We now have some information that is crucial for a utilitarian calculus. In regards to the high number of motorcyclists being hospitalised and killed, we can conclude that the majority of incidences come about as a result of riders engaging in reckless and often illegal behaviour. Riding motorcycles is an inherently dangerous activity, however, after analysing the data it is clear that it is the riders themselves who are much to blame for the exceptionally high number of motorcyclists harmed. Given the circumstances surrounding motorcycle accidents, the rate of rider casualties will not be significantly impacted by either the broad acceptance or rejection of self-driving cars, nor different approaches to programming them. Although programming self-driving cars to give priority to the safety of vulnerable road users, such as motorcyclists, would eliminate crashes involving driver error, passenger-protective self-driving cars would also put an end to such accidents, due to the fact that human driving, and thus human error, will be removed from the equation entirely. That said, if self-driving cars were programmed to prioritise those who fall under the vulnerable road user umbrella, given that many riders partake in highly irresponsible conduct, if people are sacrificed in order to spare reckless and law breaking riders, we could expect a backlash against such programming, if not self-driving cars themselves. It could also lead to more reckless behaviour from the riders, once they know that their safety will be prioritised.

Turning our attention to cyclists, we find that they account for only around three percent of all road fatalities, and fifteen percent of hospitalisations (BITRE 2015a: 1). Between 2011 and 2013, 120 cyclists were killed in Australia, with seventy-six percent of fatal crashes involving another vehicle, and twenty-four percent involving only the cyclist. Regarding the fatal multiple vehicle accidents, light vehicles such as cars were most likely to be involved in the accident, and approximately five percent of fatal accidents involved another cyclist (BITRE 2015a: 10). Of the 6,642 cyclists hospitalised in 2014, around half of such cases were due to cyclist error where they either hit a fixed object or simply lost control, while 1,414 of these cyclist accidents involved a car, although it is unclear who was at fault (BITRE 2017: 20). Important due to its emotive power, it is essential to note that the number of children killed as cyclists is extremely low, with only one child under sixteen years of age killed in 2016 (BITRE 2015a: 8). Given these numbers, cyclists do not provide a significant incentive to have self-driving cars give priority to vulnerable road users such as themselves.

When looking closely at pedestrian accidents, we find that many pedestrian casualties come about because of pedestrian behaviour. For instance, it is frequently the case that pedestrians who have been injured or killed were crossing streets or interacting with traffic in a dangerous and illegal manner (BITRE 2015b: 19). In other words, pedestrians are in large part responsible for their own harm. Researchers investigating pedestrian accidents in South Australia found that pedestrians are often struck by vehicles when they attempt to cross roads at sites which lack pedestrian crossings. Specifically, thirty-one percent of pedestrian crashes which resulted in a pedestrian being seriously injured or killed took place at intersections, sixty-five percent of which lacked signalised crossings (DPTISA 2017: 5).

Drugs and alcohol also feature in many pedestrian accidents, and there is a clear link between their consumption and pedestrian fatalities, both within Australia and internationally. Keeping in mind that the legal maximum blood alcohol level for Australian drivers is 0.05, roughly one third of all adult pedestrians killed in Australia were found to have blood alcohol levels of more than 0.08 (WHO 2013: 19). In South Australia, between 2012 and 2016, of those who were killed as pedestrians and subjected to testing, seventeen percent had a blood alcohol content greater than 0.05, and more than half of this group returned readings of 0.15 or higher. And in terms of illicit drugs, roughly twelve percent of the pedestrians were found to have used cannabis, methamphetamines, ecstasy, or some combination of the three (DPTISA 2017: 6). Similarly in New South Wales, of the ninety percent of pedestrians aged seventeen to

forty-nine killed between 2012 and 2016 and subjected to testing, thirty-nine percent were found to have blood alcohol levels of 0.05 or higher (CRSNSW 2017: 9). Across the country, it has been reported that intoxicated pedestrians are commonly hit by vehicles as they attempt crossings away from designated sites with traffic controls. And even when such pedestrians have endeavoured to cross using traffic controls such as electronic pedestrian crossings, they were rarely used correctly, a failure which may be attributable to the cognitive impairment associated with high blood alcohol levels and drug use (Holubowycz 1995).

Another factor contributing to pedestrian accidents is mobile device use. Using mobile phones and other portable devices increases the risk of being involved in traffic accidents for all road users, including pedestrians. In the United States, researchers from The American College of Surgeons (2012) found that one in five patients aged between thirteen and seventeen who had been hit by a motor vehicle were, at the time of being struck, distracted by their mobile device. That is, they were hit while focused on things such as sending messages, browsing social media, or playing music. There is even evidence suggesting that people becoming immersed in augmented reality games, such as Pokémon GO, has resulted in some pedestrians wandering directly into the paths of oncoming vehicles (Ayers et al. 2016).

Additionally, researchers (Dobson et al. 2004) exploring Australian hospitalisation records discovered that, compared to people born in Australia, those born in non-English speaking countries, or countries in which driving on the right hand side of the road is the convention, faced a much higher chance of being hospitalised or killed as a pedestrian. Arguably, this may be due to people making insufficient effort to familiarise themselves with the local conventions and language before or during their stay, or merely taking the situation of being in new and novel surroundings far too lightly. Indeed, many pedestrian casualties may be the result of non-native pedestrians lacking the skills and knowledge necessary in order to act as safe and responsible pedestrians in a foreign country.

Examining the age of pedestrians involved in accidents uncovers further important findings. According to the International Transport Forum (ITF 2012: 38), senior pedestrians (those aged over sixty-five) around the world are the most at risk group when it comes to being involved in traffic accidents. Although seniors constitute only thirteen to twenty percent of the population of OECD (The Organisation for Economic Co-operation and Development) member countries such as Canada and Australia, they account for more than fifty percent of

all pedestrians killed in traffic accidents. Oxley and colleagues have proposed that the over-representation of seniors amongst pedestrian casualties may, in part, be due to age-related cognitive decline, which renders the act of safely navigating the complexities of traffic, such as estimating the speed and distance of moving vehicles, highly difficult (Oxley et al. 2005: 962). If it is the case that a good deal of seniors are harmed in traffic accidents because of cognitive decline, we would have reason to believe that many are also unable to safely drive – that is, if they have not already been prohibited from driving for failing to pass the additional testing which senior Australians are required to undergo on a regular basis (TRMSNSW 2018). Under such circumstances, the proliferation of self-driving cars has great potential to reduce the number of pedestrians harmed, by providing seniors with a safe transport alternative that awards them greater freedom of movement in spite of the cognitive or physical barriers that may impede their safe walking or driving.

Given what I have previously said about the emotional salience attached to the death of children and the public acceptance of self-driving cars, the statistics regarding child pedestrian casualties cannot be overlooked. When we explore the data on accidents in Australia involving children up to sixteen years of age, we find that far more children die as vehicle occupants compared to children who die as pedestrians, with thirty-six children killed as vehicle occupants, and twelve killed as pedestrians in 2016 (BITRE 2017: 8). In the matter of hospitalisations, of the 2,562 pedestrians hospitalised in 2014, 398 were children, which is less than half the amount of children who were injured enough to be hospitalised as car occupants (BITRE 2017: 16).

After examining the details of pedestrian accidents, we are now in a position to reflect on self-driving car programming in relation to them. As the findings indicated, many pedestrian accidents are the result of pedestrians engaging in careless and illegal behaviour. Thus, self-driving cars which would sacrifice passengers to protect pedestrians who are getting into harms way because they are drunk, on drugs, chasing virtual characters on their mobile phones, or too impatient to cross using signalised crossing sites, etcetera, are sure to cause outrage if we see such behaviour resulting in the deaths of law abiding self-driving car passengers, especially children. This provides strong reason to opt for passenger-protective self-driving cars when looking at pedestrian incidents.

Self-driving cars, regardless of their programming, could bring down pedestrian casualties

significantly, if only by reducing the number of senior pedestrians. For instance, if seniors who are unable to drive due to cognitive or physical decline are suddenly able to utilise private or public self-driving cars, we could see a drop in the both the total number of, and amount of time, that such high-risk individuals are acting as pedestrians. If we really are motivated to limit their harm, advocating for dedicated or specialised senior self-driving cars that are provisioned to facilitate the needs of the elderly, is arguably a solution that not only benefits seniors, but also further promotes self-driving cars and reduces programming concerns. If seniors are catered for and we see a related decline in pedestrian fatalities, the overall drop in vulnerable road user casualties would, given its comparative weight against passenger casualties, produce even more reason to opt for passenger-protective programming. Otherwise, although elderly pedestrians, in instances of cognitive impairment, are unlikely to be attributed with moral responsibility in the same way fully capable motorcyclists are, we could still expect a backlash against programming that prioritised vulnerable road user safety. Especially if car occupants are killed because of the missteps of confused or disorientated senior citizens, whose engagement with modern traffic is likely to be judged as inappropriate.

With regards to children, we have reasons to favour programming self-driving cars to protect their passengers. Far more children are injured and killed as vehicle occupants than as pedestrians, thus, passenger-protective cars have greater potential to reduce the number of children harmed whilst travelling by car. Furthermore, if one accepts that less children injured and killed is preferable for the public, who have both the consumer and political power to block the establishment of a future with self-driving cars as the predominant vehicle on the road, then passenger-protective programming is more likely to facilitate the wide-spread acceptance of self-driving cars, an outcome which will decrease the number of children pedestrians harmed by driver error.

Working off the common view that the wide-spread implementation of self-driving cars will reduce current traffic accidents by around ninety percent (some say more, some less), using the data presented above, we can now get an idea of the impact self-driving cars could have on the number of people currently hospitalised and killed in traffic accidents per year. If self-driving cars were to replace human driven vehicles, it is estimated that the lives of around 750 vehicle occupants and approximately 163 pedestrian lives would be saved each year. In terms of serious injury requiring hospitalisation, we could expect up 15,000 fewer vehicle occupant hospitalisations, and a decrease of around 2305 for pedestrian hospitalisations. Thus, we are

looking at self-driving cars saving roughly 900 lives every year, and sparing something in the vicinity of 17,000 people from serious injury.

One final, and highly important piece of information regarding traffic accident statistics comes from the average number of pedestrians involved in single car/pedestrians traffic accidents. In many of the hypothetical scenarios in experimental ethics, whether they be trolley problems or dilemmas specifically involving self-driving cars, the number of people involved varies significantly so as to alter the weight between the different groups one is forced to choose between saving. For instance, trolley problems often have us choosing between five workers engaged in track repairs on one line, and a solo worker on a side-track. Or, we may be confronted with a problem where a self-driving car carrying a single passenger is confronted with several pedestrians. Such dilemmas are interesting in themselves, and they provide an opportunity for understanding our moral decision making. However, with the applied ethics of self-driving cars, we need to make decisions based on real-world data, not hypothetical cases: public policy makers cannot operate on a case by case basis, rather, they need to make decisions based on common circumstances and generalities (Goodin 1995: 69). That said, in regards to the number of pedestrians killed per fatal crash, the averages tell us that we are not dealing with scenarios like those presented in hypothetical dilemmas. Instead, there is typically one pedestrian killed per fatal pedestrian crash in Australia. Specifically, between 2007 and 2011 the average number of pedestrians killed per fatal crash was 1.01, and between 2012 and 2016, the number was 1.02 (BITRE 2017: 34). Consequently, when it comes to deciding how utilitarian cars should be programmed, we should not give weight to imagined cases in which self-driving cars are faced with large groups of child-pedestrians, for example. Rather, we should take the facts for what they are, and work on the finding that fatal accidents involving pedestrians typically involve only one pedestrian death.

Objections

Although I have touched upon and answered some general criticisms of utilitarianism, it is foreseeable that some will find the details of applying utilitarianism to self-driving cars problematic. That being the case, let us now turn to a few specific objections that are likely to arise. First, some will undoubtedly object that such an approach requires that we compare harms. But more than that, it implies that it is acceptable to do so, and that it might be worthwhile to allow the deaths of some road users within traffic accident scenarios so as to

bolster the consumer appeal of self-driving cars and reap numerous benefits. Such critics might argue that some harms, particularly death, are incommensurable. That is, the harm of death cannot be compensated for by an aggregation of outcomes of any other kind, however welcome.

In his paper *Comparing Harms: Headaches and Human Lives* (1997), Norcross grapples with the issue of comparing harms when working on moral dilemmas, and does so from a consequentialist point of view. He illustrates a type of thought experiment that often unsettles moral philosophers using the following example, which he refers to as 'life for headaches': “a vast number of people are experiencing fairly minor headaches, which will continue unabated for another hour, unless an innocent person is killed, in which case they will cease immediately. There is no other way to avoid the headaches. Can we permissibly kill that innocent person in order to avoid the vast number of headaches?” (Norcross 1997: 59).

Norcross considers several objections to the position that advocates that it would be permissible to kill an innocent individual in order to benefit others. For instance, the 'incomparability' objection holds that you cannot compare the premature death of an innocent individual with headaches. However, Norcross claims that we can dismiss this objection, because when an individual makes this type of assertion they do so because they consider killing someone to be *worse* than allowing headaches to continue, not because the two are incomparable (Norcross 1997: 60). Another objection states that the loss of innocent life is always worse than any number of other minor sufferings, and is therefore, impermissible (Norcross 1997: 61). But again, Norcross rejects this objection by introducing what he terms 'lives for convenience', which refers to most consequentialist's, and arguably the general public's, view on traffic regulations. That is, given the high probability that road deaths are positively correlated with speed limits, we could therefore reduce the road toll by simply imposing a lower mandatory speed limit of forty kilometres per hour. We do not, however, because it would be a terrible inconvenience for road users. Thus, we willingly accept deaths for the sake of convenience, which, Norcross asserts, is not significantly different, morally speaking, from his 'life for headaches' scenario (1997: 159).

The significance of Norcross' paper for the ethics of self-driving car algorithms stems from the fact that harm will never be entirely avoidable, and that we will inevitably have to face scenarios in which the options a self-driving car comes up against *all* involve causing harm to

innocent individuals. As such, we will have to grapple with the issue of comparing harms in order to decide how self-driving cars should be programmed. Although such dilemmas have no 'good' outcomes, it seems that most people would agree that some outcomes are far worse than others, and should be avoided. I argue that these harms can and should be compared to each other, in order to be able to make informed decisions. This applies even in the case of death. And given that we are willing to compare and accept some injuries and deaths for the sake of convenience in terms of speed limits, we can compare, and should accept, some harms within traffic accident scenarios in order to reap the far more significant benefits that the wide-spread implementation of self-driving cars can bring about.

An additional criticism of my approach could focus on the fact that it does not allow for any personal choice in the programming of privately owned self-driving cars, and thus is an infringement of personal freedom and autonomy. I have argued that the topic of self-driving cars and their programming needs to be dealt with from a public policy perspective, in which governments or public institutions will create rules that will apply to all vehicles. Some, such as those of a more libertarian mindset, would argue that individuals should have a say on the particular programming of their private vehicle, as it has the potential to have a major impact on their lives.

In their paper *Autonomous Cars: In Favor of a Mandatory Ethics Setting*, Gogoll and Müller (2016) deal with the issue of whether self-driving cars should be programmed with *mandatory ethics settings* (MES) that would be regulated by a third party, or whether occupants and owners of self-driving cars should be able to select *personal ethics settings* (PES), depending on their own normative persuasions. Ultimately, the authors argue that MES are a better option, for both selfish and altruistic reasons (Gogoll & Müller 2016: 681).

Gogoll and Müller note similarities between the well known thought experiments often used by moral philosophers, trolley problems, and the dilemmas that are inherent to self-driving cars, specifically, scenarios in which harm cannot be completely avoided. However, the article highlights an important distinction between self-driving car ethics and trolley problems. Namely, with trolley dilemmas, we are imagining situations where people have to act on the spot, so to speak, under time pressure with limited opportunity to gather all the relevant information. In such instances, the agent involved is unable to form a deliberative judgement, which, in turn, means that we are only able to assign responsibility to that agent in a very

weak sense, if at all. On the other hand, with self-driving cars, the agent deciding on an algorithm needs to explicitly set the rules for the car's actions, that is, the specific ethical settings, long before the car has a chance to be involved in an accident. In this regard, the individual, or regulatory body, that implements an algorithm, will be held accountable in a much stronger sense (Gogoll & Müller 2016: 683).

According to Gogoll and Müller, in modern societies, disagreement about ethical problems often results in a type of moral partitioning, where individuals are able to live in accordance with their own normative ideals. However, in the case of self-driving cars, they assert that such an approach would lead to an abundance of self-interested individuals preferring PES which would protect them as self-driving car occupants at all costs. Not to mention that some people may be inclined to go so far as to see their own cars spared damage, regardless of how minor, no matter what. In other words, they may choose a PES that directs their car to act in a manner that results in the serious injury or death of another road user, in order to prevent it from merely being scratched. Moreover, even if others adopted altruistic or utilitarian PES that would permit their vehicles to risk their safety and the physical integrity of their vehicles in some instances, eventually, such individuals would be 'crowded out', leading to the overwhelming adoption of selfish PES. Consequently, we would be left with a prisoner's dilemma situation, with all agents choosing the option that, at least initially, is in one's own best interest but which lowers utility relative to alternatives. Nonetheless, the authors state that the only way to avoid ending up in the prisoner's dilemma scenario, where everybody is ultimately worse off, is to ensure that self-driving cars are programmed with a mandatory algorithm that is enforced by a third party (Gogoll & Müller 2016: 698).

Conclusion

The ethics of self-driving cars is a topic that is starting to receive a lot of attention, and for good reason. As outlined, our current use of human driven vehicles comes with serious costs, not only in terms of lives lost or severely harmed within traffic accidents, but also in regards to public costs, both local and global. Given that the majority of traffic accidents are caused by human error, if conventional vehicles become displaced by self-driving cars, we stand to see a great reduction in harm and an increase in the public's welfare. However, self-driving cars will not eliminate all accidents. In fact, it is inevitable that self-driving cars will face accident scenarios in which harm cannot be avoided, and those involved in their programming

will be required to make decisions regarding prioritising one type of road user over another.

Because of the circumstances surrounding self-driving car programming in the Australian context, the issue is essentially an applied ethics problem in the realm of public policy. I argued that when tackling self-driving car programming, we ought to adopt a utilitarian framework. Drawing on the work of utilitarian philosophers, particularly Goodin and Singer, I outlined utilitarianism, explained why it is so well suited to the problem at hand, and addressed common concerns people have about the theory generally, as well as potential criticisms related to its application to self-driving car programming. Ultimately, I argued for a rule-utilitarian approach that is interested in producing happiness. In other words, hedonistic rule-utilitarianism.

A key aspect of taking a utilitarian approach is that it requires a thorough understanding of facts. In order to know what we ought to do, we must discern how things currently are, how our choices could impact the future states of things, who it is we are dealing with and how they are likely to be affected, which includes their prospective responses to changes in their circumstances as a result of applying self-driving car policies. Keeping in mind the limits of this thesis, I focused on key areas in order to achieve this. Namely, attention was drawn to the costs of traffic accidents, specifically, the number of people killed and injured in traffic accidents within Australia, and the impact that these accidents have on Australia's economy. Furthermore, discussion was directed to the broader consequences that traffic accidents and human driven vehicles have in the field of medicine, on land use, as well as on climate change through energy consumption and emission production. In turn, the impact that the wide-spread adoption of self-driving cars is predicted to have in all these areas was then highlighted.

As was pointed out, far more people are killed as car passengers compared to vulnerable road users. What is more, far more children are killed as car occupants rather than as vulnerable road users, which is important to remember given that we should expect the public to have particularly strong reactions to the deaths of children. By contrast, more vulnerable road users than passengers are injured in accidents. However, it was shown that a significant number of such accidents were the fault of the vulnerable road user, and often occurred without the involvement of another vehicle. For instance, many motorcycle and bicycle accidents occur due to rider error, simply through losing control or failing to avoid fixed objects.

In terms of programming, significantly more people could be spared harm if self-driving cars are programmed to give priority to the safety of their passengers. Even though some additional number of vulnerable road users may be harmed under such circumstances, they would not be numerous enough to count against adopting such programming. On the other hand, there are strong reasons against giving priority to vulnerable road users. Aside from the fact that far fewer are killed, and that a large portion of vulnerable road user casualties, due to the circumstances of their accidents, will not be avoided, no matter how self-driving cars are programmed, psychological factors provide serious reasons to opt against prioritising vulnerable road users. Specifically, such factors could work to block the broad public acceptance of self-driving cars, thus ending the chance of a number of weighty benefits coming to fruition.

To be able to reap the benefits that are predicted to come with the wide-spread adoption of self-driving cars, the new vehicles must receive the support of the public. Data coming out of experimental philosophy, coupled with an awareness of our evolved psychology, provided insight into how we should expect people to react to various approaches to programming self-driving cars. Subsequently, it was possible to reflect on the circumstances of traffic accident data, so as to envisage the public's reaction to giving priority to the safety of one type of road user over others. Upon understanding our typical psychological make-up, some programming decisions are likely to cause public outrage and have potential to lead to the broad rejection of self-driving cars. For instance, many motorcyclist casualties come about because of reckless and illegal behaviour on the part of the rider. Motorcycle accidents often involve riders speeding, being under the influence of drugs and alcohol, and not being qualified to ride a motorcycle. Similarly, pedestrian casualties have routinely been found to have consumed alcohol, or were struck by a vehicle whilst crossing the street outside of a designated area. Given people tend to possess deep-seated desires to protect the lives of their family members (who they often share vehicles with), a utilitarian would be further inclined to avoid programming that expected people to adopt a new technology which would kill them or their loved ones under some circumstances, especially when the vulnerable road user is often at fault.

Be that as it may, accounting for our psychological tendencies should not be interpreted as falling into the naturalistic fallacy. This thesis does not defend the claim that our evolved

tendencies are the hallmark of morality. On the contrary, utilitarianism is a highly rationalist moral theory that advocates a dispassionate and 'cold' assessment of consequences as a decision making process. We should, according to utilitarianism, take 'the point of view of the universe' and be as impartial as possible. However, in order to impartially assess the consequences of our decisions, we need to take into account the type of creatures that humans are. Hence, understanding the psychological lives of people is a prerequisite in order to effectively maximize their happiness.

In light of all this, I assert that a utilitarian approach to the programming of self-driving cars finds that they ought to be programmed with passenger-protective algorithms. That is, self-driving cars should be programmed to prioritise the safety of their passengers and protect them from harm, even if doing so means sacrificing a greater number of vulnerable road users in individual accident scenarios. We have enough reason to hold that people will be unwilling to purchase or utilise a self-driving car that is operating under a program that would kill them and their families if it comes down to choosing between their lives or the lives of a greater number of vulnerable road users. Not to mention that they would also object to legislation which enforced such programming. Programming self-driving cars to be passenger-protective will not only result in fewer road casualties overall, but will free up valuable public resources, in addition to having serious impacts on much broader issues, such as land use and climate change. With all this in mind, if we compare a world in which self-driving cars are programmed to cause the least amount of harm within accident scenarios (the narrow view of utilitarianism), with a world where passengers are given priority by self-driving car programming, I contend that passenger-protective programming will produce far better consequences, and as such, that is how a utilitarian self-driving car ought to behave.

References

American College of Surgeons (2012). 'Pedestrian Accidents are More Severe for Seniors and More Preventable for Young People: Trauma Surgeons Examine Injury Differences, Supervision, and Mobile Device use in Pedestrian Collisions with Motor Vehicles', https://www.facs.org/media/press-releases/cc2012/glass, Accessed 9 October 2018.

Australian Government Department of the Environment and Energy (DOEAE), 2018, 'Energy Supply, https://www.energy.gov.au/government-priorities/energy-supply, Accessed 9 October 2018.

Ayers, J. W., Leas, E. C., Dredze, M., Allem, J. P., Grabowski, J. G., & Hill, L. (2016). Pokémon GO: A New Distraction for Drivers and Pedestrians. *JAMA Internal Medicine*, 176(12), 1865-1866.

Barth, M., & Boriboonsomsin, K. (2008). Real-World Carbon Dioxide Impacts of Traffic Congestion. Transportation Research Record. *Journal of the Transportation Research Board*, (2058), 163-171.

Barth, M., Boriboonsomsin, K., & Wu, G. (2014). Vehicle Automation and its Potential Impacts on Energy and Emissions. In *Road Vehicle Automation*, 103-112. Springer.

Bauman, C.W., McGraw, A.P., Bartels, D.M. and Warren, C. (2014). Revisiting External Validity: Concerns About Trolley Problems and Other Sacrificial Dilemmas in Moral Psychology. *Social and Personality Psychology Compass*, 8(9), 536-554.

Bennett, S. (accepted for publication 2018). Are Zoos and Aquariums Justifiable? A Utilitarian Evaluation of two Prominent Arguments. *The Journal of Animal Ethics*. University of Illinois Press.

Bertalan, M. (2016). 'The Driverless Car is a Great Opportunity for Healthcare', The Medical Futurist, https://medicalfuturist.com/the-driverless-car-for-healthcare, Accessed 9 October 2018.

Bigelow, P. (2016). 'ISIS Could use a Self-Driving Car to Deliver a Bomb: Experts Warn of Security Threats in the Transportation Realm', Autoblog, https://www.autoblog.com/2016/03/15/isis-terrorists-bomb-self-driving-cars-sxsw/, Accessed 9 October 2018.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2015). Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?. arXiv preprint arXiv:1510.03346.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The Social Dilemma of Autonomous Vehicles. *Science*, 352(6293), 1573-1576.

Bradley, B. (2009). Well-Being and Death. Oxford University Press: Oxford.

Browand, F., McArthur, J., & Radovich, C. (2004). Fuel Saving Achieved in the Field Test of Two Tandem Trucks. California Partners for Advanced Transportation Technology: Berkeley.

Brown, D. (2018). 'How Self-Driving Cars or Adaptive Cruise Control Could Ease Traffic Jams', USA Today, https://www.usatoday.com/story/money/2018/07/03/self-driving-reduces-traffic-jams-study-says/741985002/, Accessed 9 October 2018.

Bureau of Infrastructure, Transport and Regional Economics (BITRE), 2015a, Australian Cycling Safety: Casualties, Crash Types and Participation Levels, Information Sheet 71, BITRE, Canberra.

Bureau of Infrastructure, Transport and Regional Economics (BITRE), 2015b, Pedestrians and Road Safety, Information Sheet 70, BITRE, Canberra.

Bureau of Infrastructure, Transport and Regional Economics (BITRE), 2017, Road trauma Australia 2016 statistical summary, BITRE, Canberra.

Burguete, P. (2016). 'Do Driverless Cars Infringe on Personal Freedom?', The Prindle Post, https://www.prindlepost.org/2016/11/driverless-cars-personal-freedom/, Accessed 9 October 2018.

Buss, D. M. (2016). Evolutionary Psychology (Fifth Edition). Routledge: New York.

Campbell, P. (2018). 'Hackers Have Self-Driving Cars in Their Headlights', Financial Times, https://www.ft.com/content/6000981a-1e03-11e8-aaca-4574d7dabfb6, Accessed 9 October 2018.

Carslake, J. (2017). 'The Most Common Causes of Car Accidents in Australia', QBE, https://www.qbe.com/au/news/the-most-common-causes-of-car-accidents-in-australia, Accessed 9 October 2018.

Centre For Road Safety Transport for NSW (CRSNSW), 2017, Pedestrian Trauma Trends Report.

Childers, J. (2018). 'How Self-Driving Cars Could Eradicate the Traffic Jam Game', Edgy Labs, https://edgylabs.com/traffic-jam-game-can-math-help-self-driving-cars-perform-better, Accessed 9 October 2018.

Crew, B. (2015). 'Driverless Cars Could Reduce Traffic Fatalities by up to 90%, Says Report', Science Alert, https://www.sciencealert.com/driverless-cars-could-reduce-traffic-fatalities-by-up-to-90-says-report, Accessed 9 October 2018.

Darwin, C. (1859). On the Origins of Species by Means of Natural Selection. Murray: London.

Dawkins, R. (2006). The Selfish Gene 30th Aniversary Edition. Oxford University Press: New York.

de Lazari-Radek, K., & Singer, P. (2014). The Point of View of the Universe: Sidgwick and Contemporary Ethics. Oxford University Press: Oxford.

de Lazari-Radek, K., & Singer, P. (2017). Utilitarianism: A Very Short Introduction. Oxford University Press: Oxford.

Dobson, A., Smith, N., McFadden, M., Walker, M., & Hollingworth, S. (2004). In Australia are People Born in Other Countries at Higher Risk of Road Trauma Than Locally Born People?. *Accident Analysis & Prevention*, 36(3), 375-381.

Federal Bureau of Investigation (FBI), 2014, 'Autonomous Cars Present Game Changing Opportunities and Threats For Law Enforcement', Strategic Perspective: Executive Analytic Report, https://info.publicintelligence.net/FBI-AutonomousVehicles.pdf, Accessed 9 October 2018.

Fleetwood, J. (2017). Public Health, Ethics, and Autonomous Vehicles. *American Journal of Public Health*, 107(4), 532-537.

Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. Oxford Review.

Garfinkel, S. (2017). 'Hackers are the Real Obstacle for Self-Driving Vehicles', MIT Technology Review, https://www.technologyreview.com/s/608618/hackers-are-the-real-obstacle-for-self-driving-vehicles/, Accessed 9 October 2018.

Gehrie, E. A., & Booth, G. S. (2017). The Impact of Driverless Cars on the US Blood Supply. T*ransfusion and Apheresis Science*, 56(2), 233.

Gogoll, J., & Müller, J. F. (2017). Autonomous Cars: In Favor of a Mandatory Ethics Setting. *Science and Engineering Ethics*, 23(3), 681-700.

Gold, N., Pulford, B.D., & Colman, A.M. (2014). The Outlandish, the Realistic, and the Real: Contextual Manipulation and Agent Role Effects in Trolley Problems. *Frontiers in Psychology*, 5, 35.

Goodin, R. E. (1995). Utilitarianism as a Public Philosophy. Cambridge University Press: Cambridge.

Government of South Australia, Department of Planning, Transport, and Infrastructure (DPTISA), 2017, 'Pedestrians Involved in Road Crashes in South Australia https://www.dpti.sa.gov.au/__data/assets/pdf_file/0020/247331/Pedestrians_-_Road_Crash_Fact_Sheet.pdf,
Accessed 9 October 2018.

Greene, J. D. (2008). 'The Secret Joke of Kant's Soul', in Walter Sinnott-Armstrong (ed.), Moral Psychology Volume 3: The Neuroscience of Morality, 35-79.

Greene, J. D. (2013). Moral Tribes: Emotion, Reason, and the Gap Between us and Them. Atlantic Books: London.

Greene, J. D. (2016). Our Driverless Dilemma. *Science*, 352(6293), 1514-1515.

Hamilton, W. D. (1964). The Genetical Theory of Social Behavior. *Journal of Theoretical Biology*, 7(1), 1-52.

Hansen, J., Sato, M., Kharecha, P., Beerling, D., Berner, R., Masson-Delmotte, V., & Zachos, J. C. (2008). Target Atmospheric CO2: Where Should Humanity Aim?. arXiv preprint arXiv:0804.1126.

Hardin, R. (1988). Morality Within the Limits of Reason. University of Chicago Press: Chicago.

Harris, J. (1975). The Survival Lottery. *Philosophy*, 50(191), 81-87.

Hilbrecht, M., Smale, B., & Mock, S. E. (2014). Highway to Health? Commute Time and Well-Being Among Canadian Adults. *World Leisure Journal*, 56(2), 151-163.

Holubowycz, O. T. (1995). 'Alcohol-Involved Pedestrians: The Australian Experience. In Proceedings of the 13th International Conference on Alcohol, Drugs and Traffic Safety, Adelaide, 13 August, http://casr.adelaide.edu.au/T95/paper/s25p2.html, Accessed 9 October 2018.

International Transport Forum (ITF), 2012, Pedestrian Safety, Urban Space and Health, OECD Publishing: Paris.

Jaworska, A., & Tannenbaum, J. (2013). 'The Grounds of Moral Status', The Stanford Encyclopedia of Philosophy (Summer 2013 Edition), http://plato.stanford.edu/archives/sum2013/entries/grounds-moral-status/, Accessed 9 October 2018.

Johnston, P., Brooks, C., & Savage, H. (2008). Fatal and Serious Road Crashes Involving Motorcyclists. Australian Government Department of Infrastructure, Transport, Regional Development and Local Planning: Canberra.

Kirkpatrick, K. (2015). The Moral Challenges of Driverless Cars. *Communications of the ACM*, 58(8), 19-20.

Knobe, J. (2007). Experimental Philosophy. *Philosophy Compass*, 2(1), 81-92.

Lafrance, A. (2015). 'Self-Driving Cars Could Save 300,000 Lives Per Decade in America', The Atlantic, https://www.theatlantic.com/technology/archive/2015/09/self-driving-cars-could-save-300000-lives-per-decade-in-america/407956/, Accessed 9 October 2018.

Leong, J. (2018). 'Study Shows Autonomous Vehicles can Help Improve Traffic Flow', Phys Org, https://phys.org/news/2018-02-autonomous-vehicles-traffic.html, Accessed 9 October 2018.

LeVine, R. A. (1988). Human Parental Care: Universal Goals, Cultural Strategies, Individual Behavior. *New Directions for Child and Adolescent Development*, 1988(40), 3-12.

Lewis, J. W. (2015). 'A Smart Bomb in Every Garage? Driverless Cars and the Future of Terrorist Attacks', START: National Consortium for the Study of Terrorism and Responses to Terrorism, http://www.start.umd.edu/news/smart-bomb-every-garage-driverless-cars-and-future-terrorist-attacks,
Accessed 9 October 2018.

Littlefield, C. H., & Rushton, J. P. (1986). When a Child Dies: The Sociobiology of Bereavement. *Journal of Personality and Social Psychology*, 51(4), 797.

Marshall, A. (2017). 'To Save the Most Lives, Deploy (Imperfect) Self-Driving Cars ASAP', Wired, https://www.wired.com/story/self-driving-cars-rand-report/,
Accessed 9 October 2018.

Mill, J. S. (2001). Utilitarianism (Ed. Sher, G.). Hackett Publishing: Indianapolis. (Original work published 1863).

Millar, J. (2014). 'You Should Have a Say in Your Robot Car's Code of Ethics', Wired, https://www.wired.com/2014/09/set-the-ethics-robot-car/,
Accessed 9 October 2018.

Millard-Ball, A. (2018). Pedestrians, Autonomous Vehicles, and Cities. *Journal of Planning Education and Research*, 38(1), 6-12.

Mitchell, A. (2015). 'Are We Ready for Self-driving Cars?', World Economic Forum, https://www.weforum.org/agenda/2015/11/are-we-ready-for-self-driving-cars/,
Accessed 9 October 2018.

Moor, R. (2016). 'What Happens to American Myth When You Take the Driver out of it? The Self-Driving Car and the Future of the Self', New York Magazine, http://nymag.com/selectall/2016/10/is-the-self-driving-car-un-american.html,
Accessed 9 October 2018.

Norcross, A. (1997). Comparing Harms: Headaches and Human Lives. *Philosophy & Public Affairs*, 26(2), 135-167.

Nourinejad, M., Bahrami, S., & Roorda, M. J. (2018). Designing Parking Facilities for Autonomous Vehicles. *Transportation Research Part B: Methodological*, 109, 110-127.

Oxley, J. A., Ihsen, E., Fildes, B. N., Charlton, J. L., & Day, R. H. (2005). Crossing Roads Safely: An Experimental Study of Age Differences in gap Selection by Pedestrians. *Accident Analysis & Prevention*, 37(5), 962-971.

Poczter, S. L., & Jankovic, L. M. (2014). The Google Car: Driving Toward a Better Future?. *Journal of Business Case Studies* (Online), 10(1), 7.

RAC (2018). 'RAC Intellibus Trial', RAC, https://rac.com.au/about-rac/advocating-change/initiatives/automated-vehicle-program/intellibus, Accessed 9 October 2018.

Road Safety Australia (RSA), 2018, 'National Road Safety Strategy', http://roadsafety.gov.au/rsa/, Accessed 9 October 2018.

Sachdeva, S., Iliev, R., Ekhtiari, H., & Dehghani, M. (2015). The Role of Self-Sacrifice in Moral Dilemmas. *PLoS One*, 10(6), e0127409.

Sidgwick, H. (1874). The Methods of Ethics. Macmillan and Company: London.

Singer, P. (1972). Famine, Affluence, and Morality. *Philosophy & Public Affairs*, 229-243.

Singer, P. (1977). Utility and the Survival Lottery. *Philosophy*, 52(200), 218-222.

Singer, P. (2011). Practical Ethics (Third Edition). Cambridge University Press: Cambridge.

Tam, D. (2012). 'Google's Sergey Brin: You'll Ride in Robot Cars Within 5 Years', CNET, https://www.cnet.com/news/googles-sergey-brin-youll-ride-in-robot-cars-within-5-years/,
Accessed 9 October 2018.

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal*, 94(6), 1395-1415.

Tinbergen, N. (1963). On Aims and Methods of Ethology. *Zeitschrift für Tierpsychologie*, 20(4), 410-433.

Transport for NSW (TNSW), 2018, 'Autonomous Vehicle Trials', Transport for NSW, https://www.transport.nsw.gov.au/projects/programs/smart-innovation-centre/projects, Accessed 9 October 2018.

Transport Roads and Marine Services NSW (TRMSNSW), 2018, 'Older Drivers', http://www.rms.nsw.gov.au/roads/licence/older-drivers/index.html,
Accessed 9 October 2018.

Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, 46(1), 35-57.

Trivers, R. L. (1985). Social Evolution. Benjamin Cummings Publishing: California.

Višak, T. (2013). Killing Happy Animals: Explorations in Utilitarian Ethics. Palgrave MacMillan: New York.

Wächter, M. A., Faulhaber, A., Blind, F., Timm, S., Dittmer, A., Sütfeld, L. R., Stephan, A., Pipa, G., & König, P. (2017). Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for ADVs. arXiv preprint arXiv:1706.07332.

Warnke, E. (2018). 'Autonomous Terror: The Clear and Present Danger of Self-Driving Cars', Hacker Noon, https://hackernoon.com/autonomous-terror-the-clear-and-present-danger-of-self-driving-cars-ed4ed1a49247,
Accessed 9 October 2018.

World Health Organization (WHO), 2004, 'The Global Burden of Disease', http://www.who.int/healthinfo/global_burden_disease/GlobalHealthRisks_report_annex.pdf,
Accessed 9 October 2018.

World Health Organization (WHO), 2013, Pedestrian Safety: A Road Safety Manual for Decision-Makers and Practitioners. The World Health Organisation: Italy. http://apps.who.int/iris/bitstream/handle/10665/79753/9789241505352_eng.pdf;jsessionid=52D010A4DCB9361864DD74E118ED2490?sequence=1,
Accessed 9 October 2018.