Ethics of Self-Driving Cars: Safety & Liability

# Introduction and Justification

Safety dominates the self-driving car discussion, representing the area of both greatest benefit and greatest concern. Proponents of the technology expound the virtues of the autonomous vehicle (AV).

The key talking points are compelling: an ever-vigilant machine that reacts at the speed of thought makes humans seem like apes behind the wheel. For detractors, safety is the area of greatest concern.   
Deep learning models like Convoluted Neural Networks (CNNs) are particularly effective at image processing and environment modelling (Barla, 2023) that make self-driving cars possible, but they are still brittle in novel situations and how they "think” is opaque. Recent advances in neural network mapping may give more insight into AI decision-making in future, but this is still far removed from the level of control we are accustomed to having over classical safety-critical automated systems.

**SAE Five Levels of Automation**

The Society of Automated Vehicles International has defined five levels of automation, with a sixth level, ‘Zero’, for unautomated vehicles. Levels 0 through 2 are categorised by features that support the driver, such as adaptive cruise control and lane departure warning. Even if the vehicle is controlling the acceleration, breaking, or steering, this classification considers that the human driver is still driving the vehicle (SAE International, 2021).

The classification considers the remaining three levels to be automated, and that a human in the driver’s seat is not driving whilst these features are active.

At Level 3, the vehicle can drive itself, but the human driver must take over when prompted to do so. Whilst SAE International (2021) characterises this as a peer of the more automated Levels 4 and 5, colloquially, vehicles with these features are rarely considered to be ‘self-driving’ cars.

At Levels 4 and 5, vehicles can complete a trip without any driver input and are typically considered to be ‘self-driving’. Fifth Level vehicles can drive themselves regardless of limitations that may be imposed on Fourth Level vehicles such as geo-fencing, poor visibility etc.

Fifth Level may be considered the ultimate goal of self-driving vehicle development (SAE International, 2021).

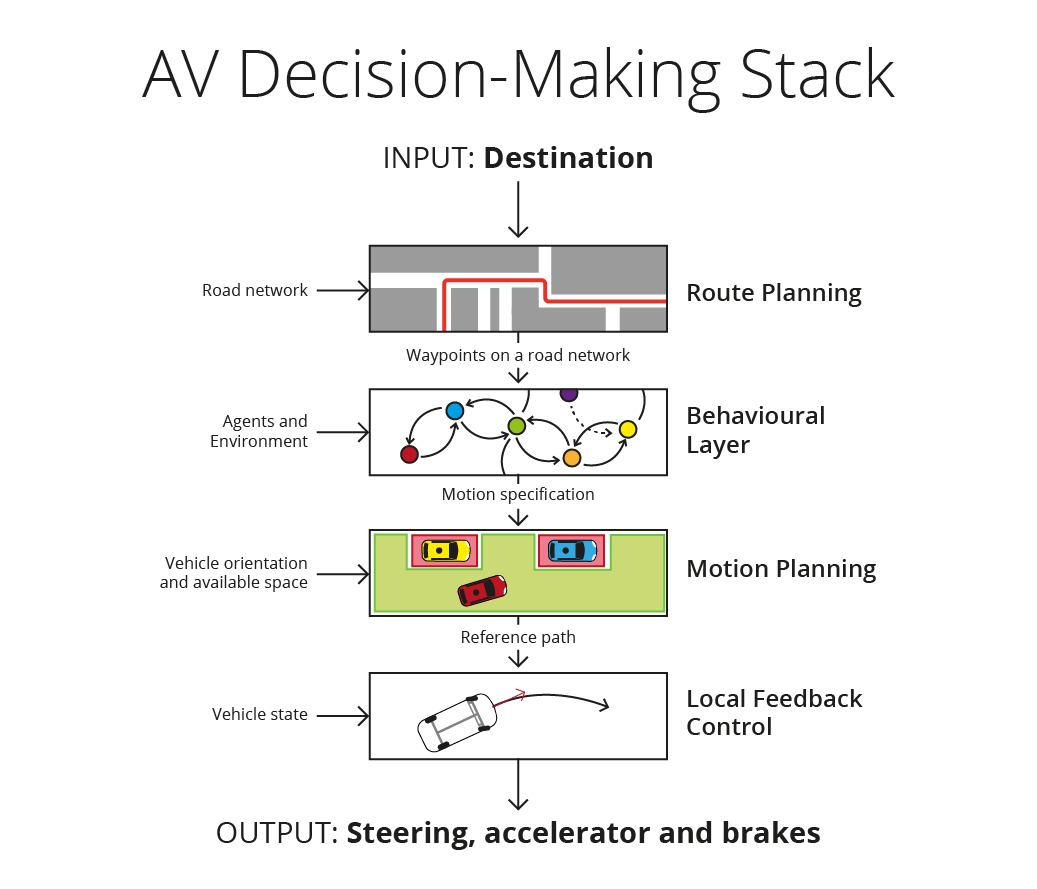
**How They Work**

The modern road environment is unpredictable, and interpreting it correctly is fundamental. To this end, AVs are equipped with RADAR (Radio Detection and Ranging) and LIDAR (Light Detection and Ranging) to map the depth of the local environment (Paden et al., 2016).

Camera arrays provide a finer picture of the near environment and use parallax to judge distance, which can be corroborated with range-finding sensors and ultrasonic sensors for judging the distance to extremely close obstacles, such as other cars and curbs. Inertial measurement units (IMUs), odometers and GPS can judge the motion, pitch, yaw and roll of the vehicle in space independently. If other more precise but more fragile sensors are rendered inoperative, the inertial sensors can function as a failsafe to bring the vehicle to a safe halt.

This data is processed in real-time by the operating system (OS) to create a map of both the static environment and dynamic objects within the environment. This map, or state, is combined with the vehicle’s state, and passed to the decision-making stack.

The first stage of this stack is route planning: the vehicle generates waypoints to its destination based on the known road layout and any perceived changes. This is then passed to a behavioural layer which “reasons about the environment and generates a motion specification to progress along the selected route.” (Paden et al., 2016)   
Using the motion specification and data about the vehicle’s orientation and local space, the motion layer then plans the vehicle’s manoeuvre as a path vector. Finally, the path vector and the vehicle’s current state are used by the local feedback control layer to generate individual inputs to the vehicle itself.



**Figure 1.** Autonomous Vehicle Decision-Making Stack (Matthew Stockdale, 2023)

All four of these layers can be implemented using classical computer science solutions, such as *A\** for route planning, and a Finite State Machine or Markov Decision Processes for behaviour (Paden et al., 2016).   
However, deep learning techniques such as Convoluted Neural Networks (CNNs) are particularly well-suited to some tasks such as image recognition and modelling spatial information. Deep learning techniques are particularly useful due to how quickly they can be trained and adjusted if relevant data is available, but some argue that they are brittle, and that not understanding how they work is inherently risky (Barla, 2023).

Classical approaches can be much more readily debugged, and their decision making can be understood by reading the code, which some consider critical for systems in which the safety of humans is involved.

**Current Status**

Since the public unveiling of the Google Self-Driving Car Project in 2010, now Waymo, investment has been thick and fast. Now, after over a decade of delays and setbacks, self-driving technology is no longer on the horizon: the push for integration is happening now.   
In August 2022, the British government unveiled their plan for a wider rollout of self-driving vehicles by 2025, backed by a £100 million investment "to support industry investment and fund research on safety developments.” (Kwarteng & Shapps, 2022)

Waymo is currently operating its ride-hailing service in Phoenix and San Francisco, with plans to move into Los Angeles. Cruise, under GM (General Motors), also offers ride-hailing in Phoenix and San Francisco. Baidu offers its 'robotaxis’ in several cities including Beijing and plans to expand to an ambitious 100 cities by 2030.   
For all three companies, safety is a key part of their branding.

New Zealand currently has such no roadmap, despite having both high car ownership and high car dependency. In the absence of government programs, usual factors such as physical isolation and a small market meant AV manufacturers had no incentive to even consider the country as a candidate testbed. However, the MOT issued a report in 2022 outlining possible impacts, as well as paths forward, and the NZTA is open to working with manufacturers interested in testing their vehicles in New Zealand.

# Opportunities

**Safety via Better Driving**

The improvements to road safety represents one of the brightest hopes of self-driving technology.   
The WHO estimated that in 2018, the number of road fatalities for that year had reached 1.35 million (World Health Organization, 2018).  
It has become the leading cause of death for persons aged 5-29, and the eighth leading cause of death for all age groups, outpacing historical killers such as disease and famine.

Research into road accidents tends to place human error as a cause more than 90% of the time. This may represent a bias against human judgement, with any failure not attributable to mechanical issues or an environmental hazard such as a landslide defaulting to ‘blame the driver’.

Improvements to road safety following the ‘Vision Zero’ principles - placing the responsibility of safety on road designersrather than road users – have seen many countries slash their road toll. Sweden leads the world in road safety, having reduced its number of road deaths by 66% between X and Y (World Health Organization, 2018).   
That Vision Zero has not reduced road fatalities to zero suggests that human error is undoubtedly still a significant contributor to road accidents.

Automated driving systems already assist in many of the sub-tasks that are necessary to drive a vehicle safely. It will likely be preferable that, at some stage, the entire responsibility of the car’s safe passage is handed over to such a system. Even before that takes place, it’s likely we will begin to see the benefits of partially automated vehicles as they lift some of the burden from drivers.

Augmenting and eventually replacing capable and responsible but fallible drivers is a worthy cause. But where self-driving cars will likely have the most to give is as an alternative to irresponsible behaviour such as driving while fatigued, distracted, intoxicated or under the effects of drugs.

Another potential use of automated driving systems is the ability to act as a co-pilot. This might be a particularly useful way to glean benefits from such systems in instances of cultures and individuals who resist giving up control to a machine.

An AI co-pilot may also be a vital training tool for inexperienced drivers and serve to resharpen the skills of trained drivers. In the case of the latter, particular care would have to be taken to ensure the system is useful without being a nuisance. Too aggressive in its instruction and it will be quickly disabled, even in cases where it might be useful.

Even the model human driver, cautious and attentive, cannot escape the decay in skill that comes with age, and like all skills, there are diminishing returns and increasing costs on further instruction to improve human driving ability.   
Inversely, new advances in AI software and hardware will likely see self-driving cars continue to improve generationally, or even over the course of the life of the vehicle via software and hardware updates.

The first laser cost as much as a small house (Downs, 1999); 40 years later, the cheapest costs less than a dollar. Therefore, it is conceivable that history may repeat itself, where sensors that are prohibitively expensive for autonomous systems now may in future be cost-effective.

All but entirely replacing the human driver may also present secondary opportunities to improve road safety. Currently, $300 million NZD is spent on policing New Zealand’s roads. In a future in which NZ’s roads are dominated by AVs designed to obey the rules of the road almost to a fault, this could be cut back significantly. Savings such as this could then be reinvested into other measures to improve road safety.

**Safety via Vehicular Design**

The most straightforward improvement AVs offer over human-driven vehicles is the lack of blind spots and the ability to observe all their surroundings at once. This is particularly significant for cyclists and motorcyclists who are much more likely to be missed by a human driver and are naturally far more fragile than their four-wheeled counterparts.

As the need to design cars around a human driver drops away, vehicular design may change to better accommodate existing or new needs, such as comfort or safety. Much of a contemporary vehicle’s surface area is glass because this is a safety necessity for a human driver. With self-driving vehicles comes the need to externalise sensors, so once the human driver is disestablished from the vehicle’s design, glass becomes a safety risk rather than a safety necessity.

Whilst still “a dream”, as Barla (2023) puts it, vehicle-to-vehicle networking is very much an avenue that will be explored once self-driving vehicles are firmly established, and as networking technologies and algorithms improve in accuracy and efficiency.

One application of inter-vehicular networking is sensor fusion. Sensor fusion is already used by cutting-edge militaries like the United States to provide a more complete picture of the battlespace to combatants, allowing them to make more informed decisions.   
Within civilian transport infrastructure, inter-vehicular networking could be used to enhance an automated vehicle’s decision-making. It could also be used to shore up an individual vehicle’s view of its surroundings if, for example, part of its sensor suite becomes incapable of providing the system with useful data in the event of damage, interference, jamming or blinding (Shalev-Shwartz et al., 2017).

Vehicle networking may pave the way for a more unified traffic management system. Such a system could influence the course of vehicles away from potentially hazardous areas, and its capabilities could even extend beyond directing traffic, such as assisting with safety and simplifying liability.   
Emergency services with access to the system could access vehicular cameras remotely to assess accident sites on or near the road to make more informed decisions when dispatching first responders.

**Safety via Predictability**

Autonomous vehicles will still share the road with human-controlled vehicles well into the foreseeable future. One advantage that may benefit the non-autonomous vehicles is predictability. The ability to anticipate other vehicles is an important skill, and one that gets easier when the vehicle being observed can be trusted to behave in a particular manner.   
With safety as a primary concern for consumers, it is reasonable to suspect, in the short term, that AVs will be overly accommodating, rather than aggressive, towards other road users.

# Risk and Challenges

**Transition Period**

The transitory period when self-driving cars are first unleashed on public roads marks one of the greatest potential risk periods. At this time, self-driving cars will be working from relatively small datasets of ‘real-world’ scenarios (Dickson, 2020) and no amount of public messaging is going to ensure that every ‘Joe Public’ is going to respond to these new cars in a sensible manner.

“It takes a long time to turn over the U.S. fleet of light-duty vehicles, with the average vehicular age currently being 11.4 years.” Therefore, “there will likely be at least a several-decade-long period during which conventional and self-driving vehicles would need to interact.” (Sivak & Schoettle, 2015)

During this transition period, it is important to acknowledge the risks of the coexistence between self-driving cars and conventional vehicles, as safety may worsen. Sivak and Schoettle (2015) state “in many current situations, interacting drivers of conventional vehicles make eye contact and proceed according to the feedback received from other drivers. Such feedback would be absent in interactions with self-driving vehicles.”

Dickson (2020) argues that we will not get Level 5 vehicles via deep learning as they must deal with human drivers still on the road, though he fails to consider that the transition period will not last forever. He primarily considers Tesla as an example for self-driving vehicles, failing to account for the other companies who are working on self-driving technology. Furthermore, Dickson takes the viewpoint that deep learning is not akin to the human mind, instead being “fundamentally flawed because it can only interpolate.”

Though this may indeed be the case, and would support Sivak and Schoettle’s absent feedback argument, the opaque nature of deep learning models cannot currently prove Dickson and those who agree with him correct.

**Deep Learning Models are Opaque in Their Decision-making**

Deep learning models, like those used in AI, are inherently opaque. Engineers may understand what the car sensed and how the car responded to a situation, but in many implementations, crucial aspects of decision-making seem to occur in a ‘black box’. (Stilgoe, 2021).

This is in sharp contrast to the state-based and deterministic automated systems to which we are accustomed. When these types of systems fail a decision-making process, the failure can be identified, fixed, and tested with a high degree of confidence. When a deep learning model fails, this is usually indicative of an alignment issue.

Alignment issues typically arise when the deep learning model learns incorrectly, such as learning the wrong goal, and thus arrives at the incorrect outcome. This is known as an “out-of-distribution generalisation” failure, where a deep learning model should perform “well on test data that is not distributed identically to the training set”, yet fails (Langosco et al., 2023).

This type of alignment issue “is a fundamental problem in machine learning” (Langosco et al., 2023) as it cannot be reasonably assumed that the model can distinguish between its “terminal goal” and its “instrumental goal” (Miles, 2021). An out-of-distribution failure could therefore cause safety issues - or even something as simple as making passengers late - with self-driving cars.

It is worth noting that whilst deep learning decision-making isn’t entirely opaque, as their learning processes can be observed, there is a lack of understanding on why they come to certain conclusions about their objectives based on their teachings, thus potentially making them dangerous.

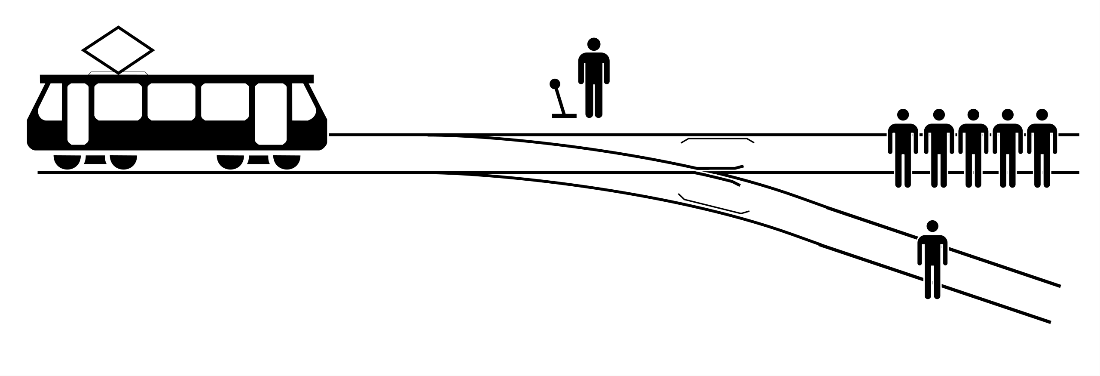
**Ethical Dilemmas & the Trolley Problem**

Addressing the problem of AI ethics will be critical for public acceptance of self-driving cars, and failure to properly align AI decision-making may have serious long-term consequences for the self-driving industry. The question of ethics is cast into its sharpest relief when introducing vehicle AI to dilemmas: scenarios in which, no matter which course the car takes, some harm is inevitable. This represents a new frontier for humanity, as never before have we tasked machines with valuing human life to this degree.

The MIT launched ‘the Moral Machine experiment’ in June 2016 with the goal of starting discussion and providing moral guidance to self-driving vehicle designers. Survey participants were provided with a birds-eye view of an ethical dilemma and tasked with picking a preferred outcome.   
The findings demonstrated a strong preference for sparing human lives over those of pets, sparing more people over fewer, sparing the young, and sparing those lawful and of higher status, such as doctors and pregnant women. (Furey & Hill, 2021)

These preferences, however, vary significantly by region. For example, countries in the ‘Eastern’ region (broadly speaking, every Eurasian country between Iran and Japan) having a preference towards the elderly. The broad array of correlations between distinct cultures suggests that the question of the machine ethics may not have a one-size-fits-all solution, despite much talk of the need for AI to reflect human values.

Furey and Hill (2020) argue that the focus on ‘trolley problem’-style dilemmas may cause more harm than good. Specifically, that by focusing on accidents that represent edge cases, the adoption of AVs may have been slowed, preventing them from saving lives attributed to one of the primary causes of road accidents: human driver error. Furthermore, Furey and Hill (2020) argue it may place unnecessary moral discomfort on the public: “[The Moral Machine] has caused the public to think they must choose between purchasing an autonomous vehicle that protects their family and one that is moral.”



**Figure 2.** “One of the dilemmas included in the trolley problem: should you pull the lever to divert the runaway trolley onto the side track?” (McGeddon & Zapyon, 2018)

Some argue that the focus on ‘trolley problem’-style dilemmas is misguided altogether. Human drivers are taught to break without swerving in the case of emergency. This procedure is not recommended merely because it is instinctive and simple, but due to the nature of vehicular physics.   
“Put in its simplest form, the problem is that swerving sufficiently to avoid an object that is within a car’s stopping distance is always a wildly risky manoeuvre compared to straight-line braking.” (Davnall, 2019)   
The self-driving trolley problem can now be seen as an additional choice between a controlled manoeuvre with known risks, and an uncontrolled manoeuvre with unknown risks.

**Need for New Infrastructure**

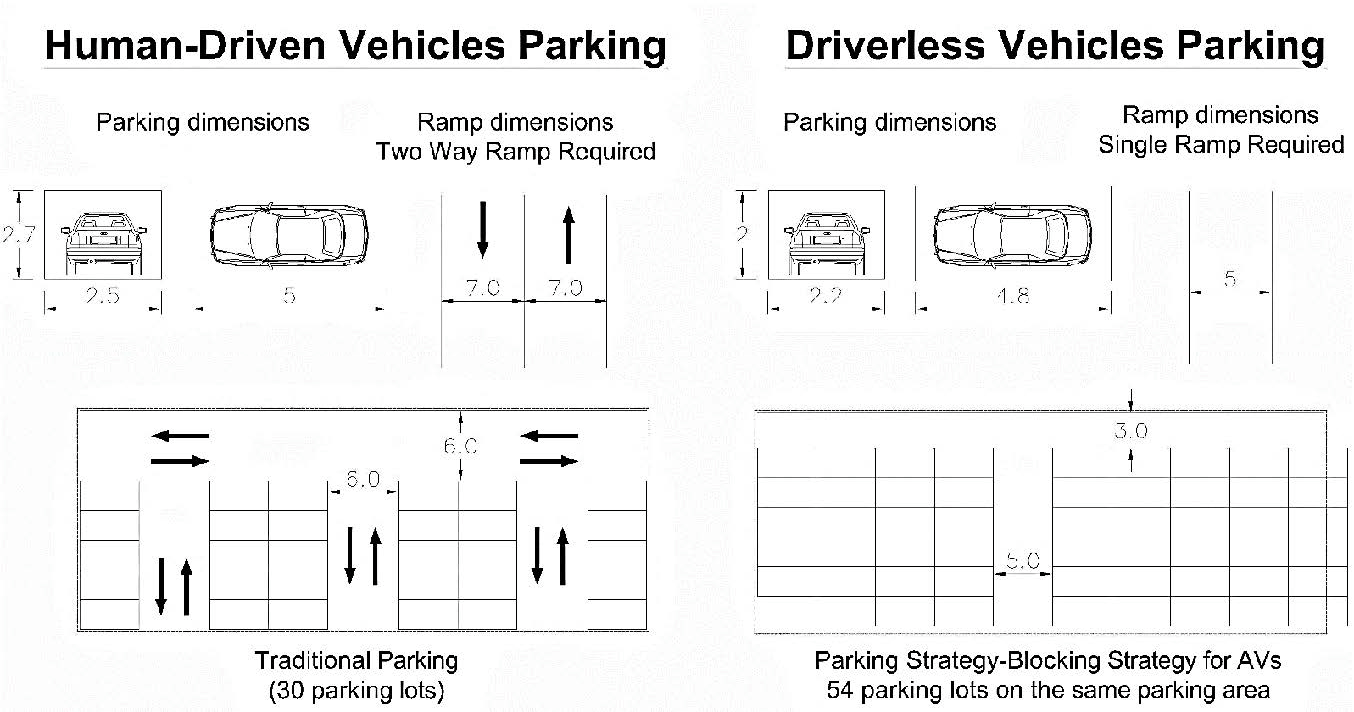
Existing vehicle infrastructure might need to be adapted for self-driving cars. As Othman (2021) states, “the human factor … will not be a concern anymore so the geometric design requirements can be relaxed.” The foremost issue with adapting infrastructure to autonomous vehicles will be the transition period, as we cannot remove all infrastructure for ‘dumb’ cars until the transition to self-driving cars is completed.

Whilst infrastructure such as roads can be streamlined during and especially after the transition period, this may increase maintenance of infrastructure. According to Othman (2021), “the decrease in the wheel wander, because of the lane-keeping system, and the increase in the lane capacity, because of the elimination of the human factor, will bring an accelerated rutting potential and will quickly deteriorate the pavement condition.”   
This could therefore drastically increase costs for infrastructure maintenance, which is already lacking for existing unautomated vehicles.

During the transition period, it might make sense to restrict self-driving cars to fixed routes and lanes, as this “allows safe operation during the early phases of driverless mobility” (Bonte, 2021).

Bonte (2021) argues that the future of cities could “favour large levels of pedestrianisation with regular traffic moved underground via tunnels and hyperloops”, although this will require a significant investment in infrastructure, which may take many years to implement even after the transition period.

“AVs can significantly reduce the number of the required parking lots … as vehicles will be serving customers at different times [and] the autonomous valet parking system will allow vehicles to park closer to each other” (Otham, 2021). This reduction of parking lots may allow us to free up space for other infrastructure such as roads or even housing.



**Figure 3.** Parking Strategy for human-driven vehicles and AVs. (**left**) Conventional parking layout; (**right**) Autonomous Vehicles parking layout. (Otham, 2021)

Despite the possible advantages, self-driving cars might have a negative impact on traffic congestion, “because AVs will motivate people to make longer trips, travel further, and make additional trips” (Otham, 2021). Autonomous vehicles therefore will possibly bring disadvantages as well as advantages.

**Need for a New Legal Framework**

Current legal framework is not suitable for handling liability in the event of an accident involving a truly autonomous vehicle. Currently, all self-driving cars on the road have a ‘steward’ or ‘safety-driver’ behind the wheel that can take control if things go wrong and is ultimately responsible for the vehicle’s safe passage. This is a temporary ‘band aid’ solution that eventually needs addressing.

Whilst it is conceivable that in the future, the AI that drive vehicles will be agents, with legal rights and responsibilities, this does not represent the immediate reality of self-driving vehicles.

With the user giving up control of the vehicle, they have also given up their responsibility for its safe use (Gurney, 2017). This stance is of absolute necessity to the adoption of AVs, as no consumer wants to purchase the ‘Car of Damocles’.

**Public Acceptance**

Having the technology of self-driving cars up to a good standard is insufficient by itself, as the public’s trust is equally important. When a technology is quite recent and seemingly beyond our control, we are naturally more risk-averse towards it.

In a study by Sivak and Schoettle (2015) on the public opinion of self-driving cars, most respondents expressed high levels of concerns about riding in a self-driving car. This is due to the fear/risk of safety issues, malfunction, or system failure. There were also concerns about self-driving cars not performing as well as human drivers.

Last year, 378 people were killed on New Zealand’s roads. This, however, is not the number to beat for autonomous vehicle engineers, as polling suggests that the public acceptance of AVs would require this number to be lowered by a factor of 10.

This risk aversion poses a risk in and of itself, as an excessively cautious approach resulting from high-profile accidents could potentially hinder the acceptance and development of self-driving cars. Such a hinderance could prevent millions from avoiding life-changing injury or even death.

# Choices

**New Legal Framework**

In creating a new legal framework capable of accommodating cases in which there is no driver to prosecute, the aerospace industry may offer a useful model to follow.

In instances of air disasters in which pilot error is ruled out, responsibility lands on those responsible for the design, manufacture, and maintenance of the systems in question.   
This is not an entirely novel approach for ground-based vehicles. The previously mentioned ‘Vision Zero’ model, which has been widely adopted in many countries, considers road designers and vehicle manufacturers to be as important to road safety as the person behind the wheel.

This may have moral shortcomings. The designers of AI based on deep learning principles cannot fully understand or predict the outcome of these systems (Miles, 2021). Risk assessment becomes progressively more difficult as the systems increase in complexity and reliability.   
To put it in plainer terms, the safer we make the self-driving car, the more difficult it is to know with certainty just how safe it is.   
As the increasing complexity of deep learning models is likely what will make the self-driving systems of tomorrow safer, it would be unreasonable to prosecute creators for failing to address fatal flaws that might take billions of hours on the road to uncover.

If serious road vehicle crashes become as infrequent as air disasters, it may even be necessary for special road crash investigation units to be set up, analogous to those that investigate air disasters.

Whilst it sounds like Asimovian science fiction, we cannot completely rule out the possibility that future AI drivers may be intelligent agents with the legal responsibilities of a person. What to do in this circumstance should remain a question for future ethicists with better understandings of these new ‘persons’, but a retreat to the familiar legal territory of the 20th century is plausible.

Unlike truly autonomous systems, Level 3 automation represents perhaps the toughest challenge for lawmakers. Treating them as if they were a Level 4 or 5 system invites abuse by users of ‘Level 3s’, who may try to pin their own mistakes on the machine.   
Without appropriate legislation, consumers may feel uncomfortable using their Level 3 system if a mistake by the system becomes their problem.   
In both cases, manufacturers are incentivised to ensure that their ‘Level 3s’ can provide impartial and indelible logs.

**Transition and Integration**

The start of the self-driving vehicle era will not mark the end of the human driver. Even places that seek to transition to a purely self-driven fleet as quickly as possible will have a transition phase. How this transition phase is approached represents one of the most significant choices that governments will have to make. Rolling out city-by-city is the approach taken by Waymo and Baidu, and has several clear advantages.

* Discussions with local governing bodies can be fronted by the company’s most senior or capable staff, increasing confidence that concerns are taken seriously.
* Marketing and awareness campaigns can be highly targeted.
* Any quirks specific to the area can be ironed out and integrated into the AVs before expanding operations further.

**Choices for New Zealand**

New Zealand, like anywhere else, needs to be ready for the self-driving car if it wants to benefit from this emerging technology, and now more than ever represents a fork in the road.

One path forward, and the path of least resistance, is to merely observe overseas developments. Superficially this would cost the government little and allow New Zealand to capitalise on the benefits whilst shouldering far fewer of the risks.   
Lawmakers would be able to follow and implement trends that work, whilst avoiding pitfalls of pioneers.   
Consumers and other road users would have confidence that when self-driving cars do reach our shores, they are less likely to suffer from the unfortunate and perhaps fatal consequences of being an early adopter.

Designing around increasingly available Level 3 self-driving cars would be the most pressing article, as they represent the most imminent change to New Zealand’s roads.

Whilst this represents the cheapest option on paper, being too slow to adopt self-driving cars might run a hidden cost. If AVs dramatically improve road safety, then everyday New Zealand is behind the curve, which could represent a genuine cost to government resources and human lives.

A more proactive path could see the government steer New Zealand towards becoming an early adopter of self-driving vehicles to address specific goals. This could be a way to shore up existing fragilities in the transport sector, such as the current shortage of bus drivers, or to reduce serious injury and death in road accidents to meet the 2030 target of a 40% reduction from 2018 levels.   
This approach would require investment and active partnership-seeking with autonomous driving system manufacturers. These partnerships would take time to develop and may not bear fruit. However, such an approach would allow New Zealand to capitalise on any benefits that manifest at the first opportunity.

The innovative path would also be the most fragile, as it would be a long-term initiative requiring constant reinvestment. A government looking to tighten its purse strings could kill the project with little to show for it.   
Putting the project on ice would not be an option. Self-driving technology is developing fast, and renewing efforts after a period of slumber may be completely unnecessary if tried and tested systems could be imported instead.   
Furthermore, stakeholders would need to be tempted in from overseas, and they would be quick to leave for areas with more developed tech sectors if the project ceased to be sustained.

# References

Barla, N. (2023, April 25). *Self-Driving Cars With Convolutional Neural Networks (CNN)*. Neptune Labs. <https://neptune.ai/blog/self-driving-cars-with-convolutional-neural-networks-cnn>

Bonte, D. (2021, April 26). *Getting cities ready for driverless vehicles*. Automotive World. <https://www.automotiveworld.com/articles/getting-cities-ready-for-driverless-vehicles/>

Davnall, R. (2019). Solving the Single‑Vehicle Self‑Driving Car Trolley Problem Using Risk Theory and Vehicle Dynamics. *Science and Engineering Ethics (2020) 26*, 431–449. <https://doi.org/10.1007/s11948-019-00102-6>

Dickson, B. (2020, July 29). *Why deep learning won’t give us level 5 self-driving cars*. TechTalks. <https://bdtechtalks.com/2020/07/29/self-driving-tesla-car-deep-learning/>

Downs, P. (1999, May). *The History of Lasers in Construction*. Construction Dimensions.

Furey, H., & Hill, S. (2021). MIT’s moral machine project is a psychological roadblock to self-driving cars. *AI Ethics 1*, 151–155. <https://doi.org/10.1007/s43681-020-00018-z>

Gurney, J. K. (2017). Imputing Driverhood: Applying a Reasonable Driver Standard to Accidents Caused by Autonomous Vehicles. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press. <https://doi.org/10.1093/oso/9780190652951.003.0004>

Holstein, T. Dodig-Crnkovic, G., & Pelliccione, P. (2018). Ethical and Social Aspects of Self-Driving Cars. *arXiv, 1802*(04103), 2-7. <https://doi.org/10.48550/arXiv.1802.04103>

Karnouskos, S. (2021). The role of utilitarianism, self-safety, and technology in the acceptance of self-driving cars. *Cogn Tech Work*, *23,* 659–667. <https://doi.org/10.1007/s10111-020-00649-6>

Kwarteng, K., & Shapps, G. (2022, August 19). *Self-driving revolution to boost economy and improve road safety*. Department for Transport; Department for Business; Energy & Industrial Strategy; Centre for Connected and Autonomous Vehicles; Centre for Data Ethics and Innovation. <https://www.gov.uk/government/news/self-driving-revolution-to-boost-economy-and-improve-road-safety>

Langosco, L., Koch, J., Sharkey, L., Pfau, J., Orseau, L., & Krueger, D. (2023). Goal Misgeneralization in Deep Reinforcement Learning. *arXiv,* 2105(14111), 1-9. <https://doi.org/10.48550/arXiv.2105.14111>

McGeddon & Zapyon. (2018, March 6). *Trolley Problem.svg*. Wikimedia Commons. <https://commons.wikimedia.org/wiki/File:Trolley_Problem.svg>

Miles, R. (2021, October 11). *We Were Right! Real Inner Misalignment* [Video]. YouTube. <https://www.youtube.com/watch?v=zkbPdEHEyEI>

Othman, K. (2021). Impact of Autonomous Vehicles on the Physical Infrastructure: Changes and Challenges. *Designs, 5*(3), 40. <https://doi.org/10.3390/designs5030040>

Paden, B., Čáp, M., Yong, S. Z., Yershov, D., & Frazzoli, E. (2016, March). A Survey of Motion Planning and Control Techniques for Self-Driving Urban Vehicles. *IEEE Transactions on Intelligent Vehicles, 1*(1), 33-55. <https://doi.org/10.1109/TIV.2016.2578706>

SAE International. (2021, May 3). *SAE Levels of Driving Automation™ Refined for Clarity and International Audience*. SAE International. <https://www.sae.org/blog/sae-j3016-update>

Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a Formal Model of Safe and Scalable Self-driving Cars. *arXiv, 1708*(06374), 9. <https://doi.org/10.48550/arXiv.1708.06374>

Sivak, M., & Schoettle, B. (2015). Road Safety with Self-Driving Vehicles: General Limitations and Road Sharing with Conventional Vehicles. *UMTRI, 2015*(2), 9.

Stilgoe, J. (2021). How can we know a self‑driving car is safe? *Ethics and Information Technology (2021) 23*, 635–647. <https://doi.org/10.1007/s10676-021-09602-1>

World Health Organization. (2018). *Global status report on road safety 2018*. World Health Organization.