Autonomous Vehicles: The Viability of Moral and Legal Agency

Lukas J. Severinghaus

Senior Thesis

May 1, 2020

Abstract:

Autonomous vehicles are being integrated into the fabric of daily life. The concept of computerized transportation promises increased safety, though at the same time exposing moral and legal issues around their implementation. Who is ultimately morally responsible for harm caused by an autonomous vehicle? Who is legally liable? Can the vehicle itself take moral or legal responsibility? This thesis proposes solutions to some of these challenges through the lens of agency status, a concept heretofore not thoroughly considered, enabling a more proactive approach to the adoption of this technology. Legal agency solves the current manufacturer legal burden and subtly modifies the legal system to accommodate self-learning systems. Moral agency complements legal agency by filling the gaps and addressing more nuanced issues of responsibility that would otherwise be left untouched. Ultimately, agency status is a vital first step to bridging the gap between rapidly advancing technology and the legal and moral societal frameworks.

Autonomous Vehicles: the Viability of Moral and Legal Agency

36,560 people died in vehicle crashes in the US in 2018 (National Highway Traffic Safety Administration, 2019). Over 100 per day, a death every 14.3 minutes, every hour, every day, every week, every year. In light of this sorrowing statistic, the vision of autonomous cars, promoted by tech magnates, seems a glimmer of hope for the future. But the fact still remains, a car is a multi-thousand pound piece of metal hurtling down the road at hundreds of feet per second. No amount of autonomy can change the facts of physics. Self-driving cars are proclaimed to save lives (Plungis, 2017), but at the same, though they may reduce crashes due to human error, they are still not perfect, and they will crash, the difference being, rather than a human killing another human, a computer is killing a human. It is not a question of if or when they will kill, they already have. On March 18, 2018, an autonomous vehicle operated by the ride sharing company Uber struck a pedestrian, who later died from her injuries (National Transportation Safety Board, n.d.). Anytime the cessation of human life is even a remote possibility, an in-depth investigation of the moral and legal issues is warranted. That is the purpose of this thesis, to examine the viability of moral and legal agency with regard to autonomous vehicles, an issue that heretofore has not been clearly researched. But first, a thorough knowledge of the background issues is imperative.

Purpose

It is vital to understand the end goal, or goals of this inquiry. While developing multiple small goals can be beneficial in some situations, within the scope of this document, it is more advantageous to specify one main target to which everything else can be measured. This not only unifies the goal but also frees the research to follow any path to the goal, rather than being constricted by multiple checkpoints.

Given this, what is the end goal? In simple terms, the end goal is to develop a concise framework for the discussion, evaluation, and integration of autonomous vehicles in our society, through the lens of the moral and legal sectors. To expand upon this, two areas bear coverage within each frame: the current problems and the foundations for future development.

On the legal front, our current judicial system is not built to handle the actions of a being that cannot ultimately be traced back to a human ("OPEN LETTER," n.d.). In short, we cannot currently ascribe accountability to autonomous vehicles. Our current legal model needs to incorporate liability for autonomous vehicles, as the vehicles are making decisions through algorithmic models outside of human prediction.

On the moral front, conventional wisdom does not accept non-sentient beings as capable of moral decision making. Therefore, one must ask, can a non-sentient being, in this case, an autonomous vehicle, become a moral agent? Principally, this matters as the vehicle is making its own decisions, following human guidance, but operating outside of direct human control. What standards are the vehicles using to make autonomous decisions? How will they handle moral dilemmas? While there is no one clear answer to every one of these questions, the need for change is evident, and the possible responses and solutions will be discussed in this paper.

Ultimately, establishing an end goal directs the course of inquiry and unites all discussions towards a common destination, an invaluable asset in such a complicated field.

What is an autonomous vehicle?

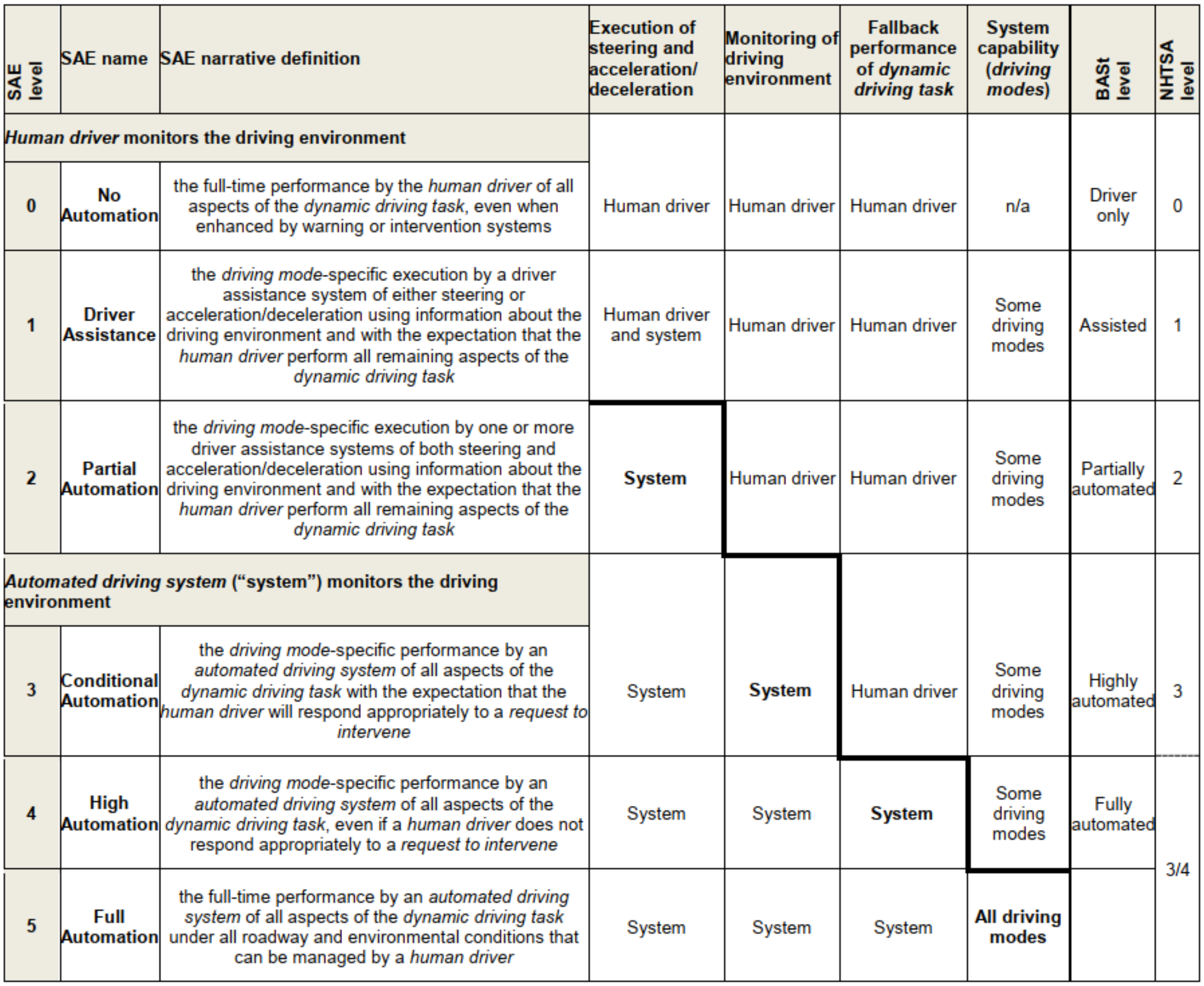
The definition of this term has the most significant impact not only on the scope of the issue but on the outcomes, both intended and unintended, that result from the discussion. Regardless of the impact, why does one need to agree upon a concise definition? Two primary benefits result from a well-scoped definition, first, enhanced clarity, and second, expanded detail.

First, a concise definition yields enhanced clarity. Especially in the fields of automobiles, thanks to the breadth of vehicle types, and as more informal and cultural language conventions creep into our vocabulary, having an established definition ensures clear communication of the scope of the meaning. Developing a formal definition for "autonomous vehicle" ensures that societal references do not adversely affect the scope of the research.

Second, a concise definition enables expanded detail. Simply, the breadth of the topic is inversely related to the depth of detail. This realization yields two conclusions: A broad view cannot show the details of specific parts, and second, one unified solution for the whole cannot prove useful for every part. If this paper were instead focused on the moral and legal status of robots in general, research would have to be conducted on every type of robot, from military robots to healthcare robots to autonomous robots to every other type imaginable. Not only is this resource prohibitive, but the scope of research would also be broader, necessitating a relatively low level of detail and in-depth thought for each category. Following from this, secondly, any solution would inevitably either completely ignore specific sectors, or contain more exceptions than actual points of implementation. By narrowing the scope to a manageable set of relatively normal situations, a more detailed investigation can be completed without sacrificing depth, detail, or dogma.

Now to move to defining the terms, what does "autonomous" and "vehicle" actually mean, and what does "autonomous vehicle" mean?

Autonomous' root concept, autonomy, is a simple notion at its core, but in the 21st century, autonomy has been attributed to a variety of machines with various capacities. Autonomous is defined as, among other meanings: "undertaken or carried on without outside control." (*Merriam-Webster.com*, 2020a) By this definition, many devices are autonomous, such as an alarm clock that rings by itself or a robot vacuum that patrols the house every day. But what happens when a device cooperates with a human? What about an unmanned aircraft, commonly known as a drone, that flies itself, but takes human inputs at times. Many of the consumer drones on the market takeoff at the press of a button, fly to a point on a map at a tap on a screen and automatically circle the pilot in "selfie mode" ("Wow Your Friends With Your Cool Drone Selfies," 2018). How does one quantify the level of autonomy when both the device and the human share control? With the rise of self-driving vehicles, the Society for Automotive Engineers International (SAE) set out to create a standard in their Surface Vehicle Information Report J3016™, entitled "Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems." (2018) In this report, the SAE outlined six levels of automation, numbered from zero to five, with differing levels of automation, refer to Figure 1 for more detail.



SAE level	SAE name	SAE narrative definition	Execution of steering and acceleration/ deceleration	Monitoring of driving environment	Fallback performance of *dynamic driving task*	System capability (*driving modes*)	BASt level	NHTSA level
***Human driver* monitors the driving environment**								
0	**No Automation**	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a	Driver only	0
1	**Driver Assistance**	the *driving mode*-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes	Assisted	1
2	**Partial Automation**	the *driving mode*-specific execution by one or more driver assistance systems of both steering and acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes	Partially automated	2
***Automated driving system* ("system") monitors the driving environment**								
3	**Conditional Automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the *dynamic driving task* with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes	Highly automated	3
4	**High Automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes	Fully automated	3/4
5	**Full Automation**	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**		

Figure 1: Diagram of SAE automation levels (Society for Automotive Engineers International, 2018)

While current vehicles on the market use phrases like "Full self-driving," ("Autopilot and Full Self-Driving Capability", 2020) currently, vehicles like the Tesla Model 3, though marketed as "Full self-driving," realistically only have autonomy at an SAE level 2, below the threshold for autonomy (Hawkins, 2019). The driver still needs to pay attention to the road and be prepared to take control at any time.

With a general idea of the current state of automation in vehicles, it is essential to define automation. For this thesis, Autonomy is classified as a system where the computer, or machine itself, has complete control of all aspects of the vehicle, where human control is possible, but not utilized in normal operating conditions. This definition is intended to line up with SAE automation levels 3-5.

When it comes to the definition of vehicle, there is a bit more consensus, given the nature of government and the current regulations on autonomous vehicles. It is best to use the definition straight out of the US Federal Code, specifically Title 49, subtitle IV, part A, chapter 301, subchapter 1, item (a)(7), established as:

> “motor vehicle” means a vehicle driven or drawn by mechanical power and manufactured primarily for use on public streets, roads, and highways, but does not include a vehicle operated only on a rail line. (“49 U.S. Code § 30102 - Definitions,” n.d.)

Using an established definition of vehicle not only promotes clarity, it enables better integration with regulators and government entities. Ultimately, based on this terminology, and the denotation of autonomous, the scope of this argument is limited to machines that mainly operate on roadways and are primarily controlled by a non-human system, i.e., a computer.

It is vital to note the distinction between self-learning autonomous systems and static autonomous systems. Current systems, though not fully autonomous, use static decision making methods, algorithms that are programmed such that any vehicle can be given the same stimuli and make the same decision. The manufacturer tells the car what to do, and nothing the user does will change it. At this early stage of autonomy, for the most part, that is a good decision. Understandably, regulators would be hesitant to allow a vehicle on the road that has a 0.01% chance of not stopping at a stop sign, so the manufacturers hard code[1] behaviors. At the same time, given the impossibility of programming the vehicle to handle every single possible situation, it is difficult to predict the action it will take in novel circumstances. For now, the uncertainty of the situation will have to be accepted. This current course of technology puts all the onus on the manufacturer. Because the user has no input in the decision making process, if the vehicle, while operating autonomously, causes injury or death, the manufacturer will be held responsible.

In the future, autonomous cars will need to move to self-learning autonomous systems. Rather than relying upon a static relationship between observing certain cues and executing corresponding behaviors, the vehicle could adapt its response to the behaviors and patterns of the environment, driver, and surrounding vehicles. Static autonomous vehicles are a rather clear cut case for manufacturer liability, the company programs the vehicle, controls its behavior, and is the only one with input to the system. Self-learning systems not only free the vehicle to adapt to its surroundings, but shed some of the liability burden from the manufacturer, and provides an

[1] Hard code is defined as “Fix (data or parameters) in a program in such a way that they cannot be altered without modifying the program.” ("Hard-Code: Definition of Hard-Code by Lexico", n.d.)

opportunity for the owner to be a part of the decision making process. Conceivably, this would come to market as a vehicle which has a base algorithm that over time and “experience” can alter certain behaviors. Optionally, companies could allow the owner to set certain preferences to also shift the actions of the vehicle, such as the acceleration limits, the aggressiveness of driving, and avoidance of certain road conditions. Together, these all slice up the “liability pie.” At delivery, the vehicle is utilizing 100% of the manufacturer’s behaviors, and thus the manufacturer is completely liable for incidents at this stage. But perhaps a few months on, the algorithm has adapted somewhat, and the owner has established certain preferences, so the manufacturer now has 30% “algorithmic control,” the vehicle 40%, and the owner 30%. With the agent-principal relationship, the vehicle would be liable for incidents arising out of any negligent adaptation, and the owner and manufacturer would be liable as principals for any actions related to their control amounts and influence. Though the exact implementation details of this system warrant further research, this provides a glimpse into how the system could function.

Autonomous Vehicle Status

Besides establishing membership in moral and legal circles, it is vital to establish a clear position in these spheres. Without this, autonomous vehicles cannot have any meaningful status, interactions, or responsibility. There are numerous possible options for designating position to autonomous vehicles, including as an entity, agent, person, electronic person, minor, or animal. This discussion will be limited to the two leading choices: agenthood and electronic personhood. [2]

[2] For more in depth coverage of the different statuses, see Gale Cengage Learning (2011), Garner & Reavley (2011), Gifis (2010), and Hall (2004).

First, autonomous vehicles as agents. An agent has an intriguing role: “Agency: as a term of art, refers to any relationship in which one person (called an agent) acts for another (called a principal) in commercial or business transactions.” (Garner & Reavley, 2011) Where other entities are separately acting “individuals” with responsibilities, rights, and obligations, an agent acts entirely within the borders of, and almost as, the principal itself. To focus back on agency, the main benefit is that the agent is acting on behalf of the principal, while at the same time being a separate entity. Autonomous vehicles operate independently, but the actions they are undertaking are those that are directed by the manufacturer, user, and owner, such as stopping at stop signs, and heeding speed limits. While operating within these directions, the principal or principals assume liability, however, if the vehicle strays from these boundaries, it assumes liability for its actions. These innate instructions link at least some liability to a responsible party while at the same time recognizing that the agent is independent of the principal and liable to make mistakes and take liability of its own.

Second, a moment of consideration is due to the 2016 European Union proposal to grant electronic personhood to autonomous robots (European Parliament Committee on Legal Affairs, 2016). Rather than classifying robots as animals, property, or persons, the document set forth a new category, that of Electronic Persons. While the concept begins to move in the right direction, a concern for consequences remote to the core issue, combined with a lack of technological expertise at the time, resulted in a document with serious flaws, as hundreds of experts pointed out in an open letter to the European Union (“OPEN LETTER,” n.d.). Principally, not enough consideration was given to all the possibilities. Quoted directly from the letter: "The economical, legal, societal and ethical impact of AI and Robotics must be considered without haste or bias." (“OPEN LETTER,” n.d.) This is a valid point, a group of legislators cannot develop a practical, effective solution, especially in such a uniquely technical space, without extensive consultation and participation from experts in the industry. Secondarily, they asserted that the creation of electronic personhood to determine liability, in the belief that robot actions cannot be traced back to a human, is faulty. Again, quoted from the letter: "From a technical perspective, this statement offers many biases based on an overvaluation of the actual capabilities of even the most advanced robots, a superficial understanding of unpredictability and self-learning capacities and, a robot perception distorted by Science-Fiction and a few recent sensational press announcements." (“OPEN LETTER,” n.d.) While an interesting subject of research, the flaws exposed in the above letter exclude electronic personhood from consideration as the best status for autonomous vehicles.

Having examined agenthood and electronic personhood in more depth, as a potential model for autonomous vehicle status, agenthood presents the best option for three main reasons: Agents act on behalf of the principal, agents do not automatically assume rights like other statuses, and the liability for the actions of the agent rests primarily on the principal, with exceptions for negligence.

First, agents act on behalf of the principal. This is in the very definition of an agent, “one person(called an agent) acts for another (called a principal)” (Garner & Reavley, 2011) Through this, it becomes clear that though the agent is acting by itself, it is acting within the boundaries established by the principal. The agent is free to act, but only to do so within boundaries. This closely matches how autonomous vehicles act, they can make decisions that cannot always be humanly predicted, like with the use of artificial intelligence, but at the same time, there are specific rules they cannot violate.

Second, agents do not automatically assume rights like other statuses. Agenthood, in its contemporary use, is meant to build on top of an existing status, per the definition “one person(called an agent) acts for another.” (Garner & Reavley, 2011) The agent is not an existent class in and of itself; it operates on top of an already established framework. It thus does not communicate the rights and obligations that some other statuses, like persons or entities, inherently possess.

Finally, the liability for the actions of the agent rests primarily on the principal, with exceptions for negligence. In an article in the *Harvard Law Review*, the liability of principals for the actions of the agents is set out as follows: “In general, an agent negligent in the performance of his duty is liable to the principal for all damages proximately resulting from that negligence.” (1919) This is advantageous for autonomous vehicles, as the principal, either the manufacturer, user, or a combination of the two, is liable for most actions of the vehicle. However, if the vehicle, for some reason, causes a negligent outcome, some liability could be shifted to the vehicle.

In summary, agenthood acknowledges the reliance of autonomous vehicles on human background. It balances the liability of human actors with the actions of the vehicle itself, while restricting the vehicle's rights.

This paper proposes that agenthood is the best choice for the status of autonomous vehicles, but first a potential concern needs to be addressed. Can autonomous vehicles take liability for their actions? In the open letter opposing the EU Electronic Personhood Proposal ("OPEN LETTER," n.d.), the writers stated that the current state of robots, from a technical perspective, enabled every action to be traced back to a human actor. While that may have been the case at the time of writing, with the rise of artificial intelligence and machine learning systems, if not now, then in the very short term, it will be challenging to trace certain behaviors back to one responsible party, as Peter Asaro stated in his conference paper entitled, "The Liability Problem for Autonomous Artificial Agents":

> Currently, it is possible to analyze and test a learned function and determine its behavior, as with traditional engineering. But when AI systems are allowed to continue modifying their functions and learn after they are deployed, their behavior will become dependent on novel input data, which designers and users cannot predict or control. As a result, the behavior of the learned functions will, to various degrees, also be unpredictable. (Asaro, 2016)

Our contemporary legal system can only hold liable those who take part in an act. If self-learning systems are truly self-learning, and have little or no human input, the systems will need to be held accountable for their actions. Though there is no doubt of the need for autonomous vehicle liability, the actual implementation of it, along with the inevitable caveats, needs to be discussed in the context of the broader legal argument.

Ultimately, though every approach has downsides, giving autonomous vehicles agenthood is the best choice to balance the rights and liabilities of the vehicle, owner, and manufacturer.

Now, with a thorough explanation of the background details, the core argument can be stated. Autonomous vehicles should have moral and legal agency status. This can be subdivided into two parts, legal agency, and moral agency.

Legal Agency

Conventional road vehicles operate under a set of laws. Drivers have to comply with legal requirements regarding operation and safety. Autonomous vehicles will have to comply as well. However, in order to comply, there must be someone to hold accountable. A judge can not send an autonomous vehicle to jail or order it to pay a fine, it is not sentient, and it is incapable of earning money. Legal agency promises a solution to this, but before that, what is legal agency? In short, it is: “[a] consensual relationship created by contract or by law where one party, the principal, grants authority for another party, the agent, to act on behalf of and under the control of the principal to deal with a third party.” (Gale Cengage Learning, 2011) Basically, the autonomous vehicle, acting as the agent, operates within the limitations imposed by the principal, who could be the owner, manufacturer, a third party, or a mixture thereof. The critical implication of agency comes with the liability of the principal. As stated previously, generally, the principal is liable for the actions of the agent, except in cases of negligence on the part of the agent (*Harvard Law Review*, 1919). This is one of the primary benefits of agency, as responsibility can be shifted away in proportion to the independent learning capabilities of the machine.

To argue both sides, why should autonomous vehicles have legal agency? First, this status ensures that autonomous vehicles can take responsibility for their negligent actions. If an autonomous vehicle, through its self-learning capabilities, causes harm, the vehicle itself will be held responsible. This is not to say that autonomous vehicles are physically or mentally capable of defending themselves in court, a human entity will have to represent them, a point that will be shortly expanded. Second, this shift in our legal system will acknowledge the direction the technology is headed. Through altering our legal system to allow non-sentient machines to take responsibility, we will embrace the advance of this technology while ensuring all types of self-acting beings are justly held liable for their actions. Finally, as a cumulative benefit of these changes, legal agency will free innovators to develop groundbreaking technology without fear of unnecessary litigation. Currently, manufacturers stand to be sued every time one of their autonomous vehicles crashes, legal agency would help alleviate some of the litigious burdens from actions negligently taken by the autonomously operating vehicle.

At the same time, there are a few notable challenges of legal agency. Shifting accountability away from human parties could encourage individuals or companies to exploit the autonomous vehicle's learning system, or to cut corners in crucial software components while hiding that within the self-learning aspects of the vehicle. To be clear, the agency relationship only shields the principal from liability in the case of negligence of the agent; the companies will only be shielded from explicit negligent action of the vehicle. However, as with anything, it is bound to be abused, and unfortunately this can only be remedied with time and experience.

Additionally, through this concept of legal agency, responsibility would be delegated to a machine that is incapable of representing itself. While this seems like a formidable challenge, our society has already tackled a similar problem, and the same solution can be applied to this issue. The insurance industry already provides coverage for millions of drivers, and could easily cover, and represent, autonomous vehicles. At its core, the autonomous vehicle would take liability, but the insurance company would represent the vehicle and satisfy the claims of the litigants, much like our current system. The company would assume the penalties and burdens that the vehicle could not fulfill, drawing from the collected premiums of their clients. In addition to solving the liability issue, by privatizing the coverage of autonomous vehicles, the insurance companies could, and likely would, in the pursuit of financial gain, develop systems to classify the risk of autonomous vehicles, thereby incentivizing the vehicle owner, and the manufacturer of the vehicle, through reduced premiums, to maintain a higher standard of care in the design, operation, and maintenance of the vehicle, much like the automobile “ecosystem” of today.

Ultimately, there are three substantial implications for this choice. First, humans would not be liable for the negligence actions of autonomous vehicles. Second, a new private system, similar to our current automotive insurance system, would need to be developed and implemented.

Finally, and perhaps most significant, our legal system would shift to allow liability to be given to non-sentient machines.

What if autonomous vehicles were not afforded legal agency? This position acknowledges that autonomous vehicles cannot take responsibility for their actions, leaving all liability to a human. Additionally, this necessitates no change in the legal framework. Nevertheless, at the same time, the lack of legal agency could very well stifle innovation as companies could be found liable for the independent decisions of their products, thereby discouraging the use of self-learning systems. The innovation impact, though aside from the core issue, when combined with the difference in liability, shows that this course of action does not line up with the end goal, and thus is not the correct course of action.

Ultimately, autonomous vehicles are best suited to have legal agency, both to acknowledge the core technology, and to shift the legal burden away from humans. Necessitating only small modifications to our legal framework, this move is most prudent considering the advancing pace of technology and the necessity of a legal system that acknowledges non-human actors. At the same time, laws are only half of the picture. Where laws are black and white, morality is concerned with a completely different dimension. Autonomous vehicles must be moral in their actions, in addition to being legal.

Moral Agency

What is moral agency? But first, what is moral? Moral is defined as: "of or relating to principles of right and wrong in behavior." (*Merriam-Webster.com*, 2020b) This definition raises an important point regarding the phrase: "right and wrong" (*Merriam-Webster.com*, 2020b). Morality, which, paraphrased from the dictionary, is the practice of being moral (*Merriam-Webster.com*, 2020c), judges actions on a scale of rightness. Therefore, particular behaviors in certain situations can be regarded as right, for example, helping an elderly person cross the street. In contrast, others can be regarded as wrong, for example, kicking an animal. At the same time, these actions are context-dependent; the actions are not right or wrong in and of themselves. Helping an elderly person cross the street into oncoming traffic is obviously wrong, while at the same time kicking an animal to protect one's child would be regarded as right. While some would also argue that thoughts can be judged as moral, this discussion will be limited to morality insofar as it refers to actions. Stimuli, whether good or bad, prompt every action, and thus every response taken by the autonomous vehicle has a moral rightness. A search for the topic of moral reasoning in robotics yields dozens of research papers exploring the possibility of moral cognition[3]. Nevertheless, what if, instead of developing moral cognition, moral agency was fostered instead?

With moral defined, how does agency come into play? Much like legal agency, moral agency allows the autonomous vehicle to act within the bounds set by the principal, either the manufacturer, owner, user, or a combination thereof. Because the principal establishes the boundaries for the vehicle, it logically follows that the morally responsible party is the principal,

[3] Moral Cognition is defined as "the study of the brain's role in moral judgment and decision-making." ("Moral Cognition," n.d.) For more information regarding the concept, see https://ethicsunwrapped.utexas.edu/glossary/moral-cognition

unless the vehicle negligently strayed from those bounds, in which case the vehicle would be morally responsible, a concept that will be discussed shortly. The autonomous vehicle is not making its own moral decisions; it is merely following the guidance provided by the principal. In self-learning systems, this could manifest through a sequence of pre-established scenarios, where the correct outcome is communicated to the vehicle, such as “stop at stop signs” and “yield to pedestrians.” Given thousands of these scenarios, a self-learning system could develop a general idea for actions that are right and wrong, thereby mimicking the moral stance of the principal without having to engage in human-like moral cognition.

What if autonomous vehicles did not have moral agency status, and specifically lacked the user-controlled moral adaptability of the previous section? The status quo, devoid of moral agency, raises two issues. First, the issue of moral paradoxes. What action will the vehicle take in response to unavoidable casualties and other complex situations? Second, the issue of moral responsibility. Who is morally responsible for the actions of the robot?

Paradoxes are best explored in the mind. Imagine an autonomous vehicle traveling at highway speeds. A pedestrian appears directly in front of the vehicle, and the vehicle is forced to decide between running over the pedestrian or driving into a barrier, killing the human occupant of the vehicle. What will it choose? Autonomous vehicles will encounter these types of situations with unavoidable casualties. It becomes a decision of whether the life of one is more important than the life of another. In practice, developing one single answer to this problem is impossible, as a study referenced in the MIT Technology Review showed: “People are in favor of cars that sacrifice the occupant to save other lives—as long [as] they don't have to drive one themselves." ("Why Self-Driving Cars Must Be Programmed to Kill," 2015). This quote ironically illustrates not only the moral dilemma, but the inherent bias for self preservation. At the end of the day, these paradoxes are just that, paradoxes, there will never be one perfect solution for everyone. Though the moral conservation must continue, the first necessary step, through the implementation of moral agency, is to allow the owner or user to participate in the principal role of moral agency and have a hand in the moral decision-making process. Not to say that this is foolproof, it has the potential to be abused, but at the end of the day, it moves the most in the right direction.

Moral responsibility for autonomous vehicles is currently a tenuous issue. As of the time of writing, the vast majority of "autonomous vehicles" on the road are not fully autonomous but instead advanced "driver assist" technologies. Though the current principles of autonomous vehicle manufacturers are implemented in vehicles that are not fully autonomous, looking at their approaches yields a glimpse into how the vehicles will approach problems when they do become fully autonomous. A 2017 *Forbes* article covered a class-action lawsuit regarding the semi-autonomous vehicle manufacturer Tesla's use of their Autonomous Emergency Braking (AEB) system (Lin, 2017). In that article, the main legal complaint was that the AEB system, which could detect an impending collision with a frontal object and brake to avoid or reduce the damage in a collision, would not engage if the human driver was pressing the accelerator pedal, or had pressed the brake pedal (Lin, 2017). This is the heart of the dilemma. If the vehicle could have avoided the collision by overriding the human input, but did not, the vehicle caused harm through inaction, but if the actions of the human driver would have avoided the collision, but the vehicle overrode human input and caused a collision, then the vehicle caused harm through action. Obviously, these are only two of four possible outcomes, the human could have avoided the collision while the vehicle did not override the controls, or the vehicle could have avoided the collision by overriding the controls, and either outcome would have resulted in no harm, but either way, through overriding the controls, or not, there is still a possibility that a collision will occur. Tesla leadership has decided that, rather than developing a solution to the dilemma or allowing users to pick their own answer to it, they will instead inhibit the vehicle from taking control, thus taking their vehicle, and their company, out of the liability of the situation. Rather than thoughtfully using the tools at their disposal to help avoid accidents, they have decided to completely step out of the situation, making their car no better in that situation than a car with no automated features whatsoever. Alternatively, a system with a human driver entirely in control and an automated system that could take control if it was programmed to, would leave the morally responsible person of the situation indeterminable. This is the tenuous state of semi-autonomous vehicles, the same state of affairs that will help shape the fully autonomous vehicles of the future. As of now, there is no solution to autonomous moral responsibility, though the need for it is evident. Unfortunately it appears that technology has outpaced the modern advancement of morality, and that these concepts need to be developed further to find an answer. Ultimately, they will need to be expanded upon as a part of the broader discussion around morals and the definition of morality.

Moral responsibility is a vital component for the future of autonomous vehicles. Some would argue that only legal responsibility is necessary, surmising that morality cannot cause someone to change their actions, or be enforced, like our system of laws. This stance completely ignores the moral paradoxes, situations where either decision is harmful to one party but ultimately possesses the same legal implications, killing a person. Our legal system is only intended to work with laws, rules that show right and wrong; it is not equipped to deal with two acceptable choices that both have consequences.

Furthermore, some would argue that under no circumstances can autonomous vehicles exercise morality in actions. Johnson and Axinn concluded:

> Autonomous robots with no human in the loop cannot be moral actors. They lack both the imagination to conceive of the effects should the principle of their actions be made universal, as well as the free will to make the choice to follow a moral style. There is no test of morality that a robot could pass as such as only the actions resulting from moral decisions are testable. They may appear to be acting morally, as they may take the same action we would expect a moral person to take, but that does not make them moral. For these reasons, they should not be employed in situations requiring moral action. They cannot be trusted to decide on killing humans, or on attacking buildings or vehicles, they should certainly have no autonomous lethal use. (2018)

This raises a valid point, autonomous vehicles can not, and likely will never be able to possess the mental capacities to make moral decisions on their own. But conversely, the point that, as they cannot morally decide, they should be prohibited from morally acting is overly restrictive.

Just because a situation would require a moral decision does not mean that a machine should have to consider the possibilities and make a moral judgement. Moral agency allows an autonomous vehicle to take actions that are considered moral, thus allowing moral action without moral reasoning. The vehicle can use its training data of moral and immoral actions and contexts to determine the moral action for the situation. Again, this is not to say that the system will be perfect, but the alternative is that we remove automation from anywhere that there is even a remote possibility of human harm, or another situation requiring moral action. While Johnson and Axinn (2014) raises some valid points, ultimately moral agency provides the best route forward without completely devolving the pace of technology.

For these reasons, it is imperative not only that the legal responsibility and moral responsibility be developed in parallel, but that the moral behavior of autonomous vehicles base their actions off the guidance of a human as we advance into a future of vehicles that will have to make the hard choices about human life.

Conclusion

Autonomous vehicles promise a safer future for transportation. They do not get bored, tired, or lazy, unlike their human counterparts. But at the same time, they are not immune to failure, and at this stage, the stakes of failure are high. As autonomous vehicles play a greater role, it is imperative to have a firm grasp of the moral and legal issues surrounding their implementation, in order to minimize unintended consequences. Legal agency promises enhanced fairness and accountability to all parties involved, while not hampering technological innovation. Moral agency, though not yet fully developed, helps to promote these same values where the legal system falls short. Combined, these two concepts enable users, manufacturers, and regulators to move from a reactive to a proactive approach with regards to this technology. Ultimately, having a grasp of the moral and legal landscape, both the issues and the solutions, enables development and adoption of the technology to continue in a thoughtful, responsible, and proactive manner.

References

49 U.S. Code § 30102 - Definitions. (n.d.). Retrieved April 7, 2020, from https://www.law.cornell.edu/uscode/text/49/30102

Agency. Principal's Rights against Agent. Recovery for Agent's Negligence Denied Where Equivalent to an Indemnity for Principal's Own Tort. (1919). *Harvard Law Review*, 33(1), 106-107. doi:10.2307/1328093

Asaro, P. (2016). "The Liability Problem for Autonomous Artificial Agents,"AAAI Symposium on Ethical and Moral Considerations in Non-Human Agents, Stanford University, Stanford, CA, March 21-23, 2016.

autonomous. 2020a. In *Merriam-Webster.com*

Retrieved 4/7/2020 from https://www.merriam-webster.com/dictionary/autonomous

Autopilot and Full Self-Driving Capability. (2020, March 27). Retrieved April 7, 2020, from https://www.tesla.com/support/autopilot#capability-features

European Parliament Committee on Legal Affairs. (2016) DRAFT REPORT with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)) Retrieved from

https://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf?redirect

Gale Cengage Learning. (2011). Gale Encyclopedia of American Law Volume 1. Detroit.

Garner, B. A., & Reavley, T. M. (2011). Garner's dictionary of legal usage. Oxford: Oxford University Press.

Gifis, S. H. (2010). Law dictionary. Hauppauge, NY: Barron's Educational Series, Inc.

Hall, K. L. (2004). The Oxford Companion to American Law. Oxford: Oxford University Press.

Hard-Code: Definition of Hard-Code by Lexico. (n.d.). Retrieved April 24, 2020, from https://www.lexico.com/en/definition/hard-code

Hawkins, A. J. (2019, January 31). No, Elon, the Navigate on Autopilot feature is not "full self-driving". Retrieved April 7, 2020, from https://www.theverge.com/2019/1/30/18204427/tesla-autopilot-elon-musk-full-self-driving-confusion

Lin, P. (2017, April 5). Here's How Tesla Solves A Self-Driving Crash Dilemma. Retrieved April 8, 2020, from https://www.forbes.com/sites/patricklin/2017/04/05/heres-how-tesla-solves-a-self-driving-crash-dilemma/#6ac129356813

Moral Cognition. (n.d.). Retrieved April 21, 2020, from https://ethicsunwrapped.utexas.edu/glossary/moral-cognition

moral. 2020b. In *Merriam-Webster.com*

Retrieved 4/7/2020 from https://www.merriam-webster.com/dictionary/moral

morality. 2020c. In *Merriam-Webster.com*

Retrieved 4/7/2020 from https://www.merriam-webster.com/dictionary/morality

National Highway Traffic Safety Administration. (2019). 2018 Fatal Motor Vehicle Crashes: Overview (Report No. DOT HS 812 826). Retrieved April 18, 2020, from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812826

National Transportation Safety Board. (n.d.). Preliminary Report Highway (Report No. HWY18MH010). Retrieved April 18, 2020, from https://www.ntsb.gov/investigations/AccidentReports/Reports/HWY18MH010-prelim.pdf

OPEN LETTER TO THE EUROPEAN COMMISSION ARTIFICIAL INTELLIGENCE AND ROBOTICS. (n.d.). Retrieved April 8, 2020, from http://www.robotics-openletter.eu/

Plungis, J. (2017, November 7). Faster Rollout of Self-Driving Cars Would Save Lives, Study Says. Retrieved April 18, 2020, from https://www.consumerreports.org/autonomous-driving/faster-rollout-self-driving-cars-would-save-lives/

Society for Automotive Engineers International. 2018. Surface Vehicle Information Report J3016TM. Retrieved February 10, 2020, from https://www.sae.org/standards/content/j3016_201806/

Why Self-Driving Cars Must Be Programmed to Kill. (2015, October 26). Retrieved April 8, 2020, from https://www.technologyreview.com/s/542626/why-self-driving-cars-must-be-programmed-to-kill/

Wow Your Friends With Your Cool Drone Selfies. (2018, April 23). Retrieved April 7, 2020, from https://store.dji.com/guides/wow-your-friends-with-your-cool-drone-selfies/