Self-protective versus utilitarian autonomous vehicles

The following examination focuses on the moral dilemmas surrounding the comparison of self-protective and utilitarian autonomous vehicles. These vehicles can be programmed to prioritize the safety of passengers in an accident or prioritize the greater good to save more lives. The essay will explore various ethical questions such as evaluating the numbers game approach, analyzing the principles of beneficence correlated with social inequality, and interpreting the principle of autonomy in the context of autonomous vehicles. Additionally, the examination will consider the harm–benefit ratio. In the sense of methods, this analysis uses the classical issues in a new way and provides recommendations for decision-makers to consider.

Keywords: *self-driving cars, utilitarianism, self-protective, moral issues*

Author Information

Beáta Laki, University of Pécs Medical School, Department of Behavioural Sciences
https://orcid.org/0000-0002-6348-8483

How to cite this article:
Laki, Beáta. “Self-protective versus utilitarian autonomous vehicles”.
Információs Társadalom XXIII, no. 2 (2023): 83–94.
https://dx.doi.org/10.22503/inftars.XXIII.2023.2.5

1. Introduction

In my following essay I aim to explore the ethical dilemmas surrounding the comparison of self-protective and utilitarian autonomous vehicles (AVs) in a unique way. Autonomous cars can be programmed or taught to be self-protective vehicles, when the highest rule to follow is the safety of the passenger in case of an accident situation or through utilitarian approach to serve the highest good and save as many lives as possible.

There are ongoing debates about which of these we should use and choose as owners of AVs. We are faced with many arguments in favor of and against each. But at least one fact is sure: the consequence of introducing autonomous cars into global traffic saves many lives. Because of the nature of AVs, they are able to quickly calculate the best decision in an emergency situation, whereas human drivers may not be able to do so; humans are not capable of thinking through all of the possible scenarios and acting in the best way in such an urgent situation.

This raises questions about how we should choose between the two types of programmable "act" as drivers and as pedestrians? Should we merely play the numbers game or is something else essential? What is the best moral decision, if such a decision exists at all? Is postponing introducing AVs not an immoral behavior by itself?

In this essay I will circle around these kinds of question, taking into consideration, inter alia, the harm–benefit ratio, personal autonomy, the possible wider effect on the whole of society, the lives that could be saved with and without AVs, and issues of social and personal responsibility.

2. Why do we use AVs[1]?

The primary goal of using AVs is safety and crash prevention, but this has proven to be a complex issue. Before delving into the details, it is worth noting the various benefits of using these machines in everyday life. (Maurer et al. 2016)

Without exhaustive list of benefits, AVs make traveling more comfortable and smooth. They decrease traffic violations. They also provide more safety than traditional driving due to the lack of driver fatigue or impairment: passengers do not have to be afraid of the driver becoming tired, or being under the influence of alcohol or drugs. And if we compare the reaction times, we can find significant difference also: AVs have faster reaction times than humans, making them better equipped to handle unexpected situations (Braun et al. 2020). The environmental effects of AVs are also less due to the system being able to choose the driving mode that uses only as much fuel/energy as necessary; they do not waste fuel on useless fast-speed acceleration, etc.

[1] In this essay when I mention AVs I understand them to be 4th level – high automation – and higher. The 5th-level – full automation – AVs are those that do not need a human to intervene in the driving; they are capable of driving long distances without human support. They are real self-driving cars (Britton 2020).

Despite these benefits, a significant problem remains with the use of AVs: how to protect and make decisions in the event of an accident. How do we have to protect at all? And how to decide to choose one or the other participant of the accident who will be the victim if the harmful event is inevitable?

The book *Moral Machines* (Wallach and Allen 2009) concludes that as autonomous systems become more prevalent in our society, it is important to consider the ethical implications of their actions. The authors argue that current approaches to programming autonomous systems, such as rule-based systems and utility-based systems, are not sufficient for dealing with the complex moral dilemmas that autonomous systems may face. They propose the use of a "hybrid approach" that combines elements of rule-based and utility-based systems with a more comprehensive understanding of ethical principles. In summary, the book concludes that autonomous systems have the potential to greatly benefit society, but that there are also significant ethical challenges that must be addressed in order to ensure that these systems are used in a responsible and ethical manner.

It seems, then, that moral theories may not be that helpful when it comes to real-time decision procedures, since none of them can solve the situation beneficially for everyone. In the following I will analyze this issue from different perspectives and with the help of some arguments of applied ethics.

2.1. An unsolvable problem?

When creating regulations for AVs, two main interests must be taken into account. One is the interest of buyers who want an AV that protects them and saves their lives in case of an accident, while also serving their needs and making their life easier. The other is the interest of pedestrians who want to experience the benefits of AVs in terms of protection and safety. Both of these groups want the same thing: protection and safety. However, it is not always possible to save all lives in the event of an accident, which creates a difficult problem in terms of regulation. At the same time, it is important to create regulations that are acceptable to all parties, regardless of whether they are AV owners or pedestrians.

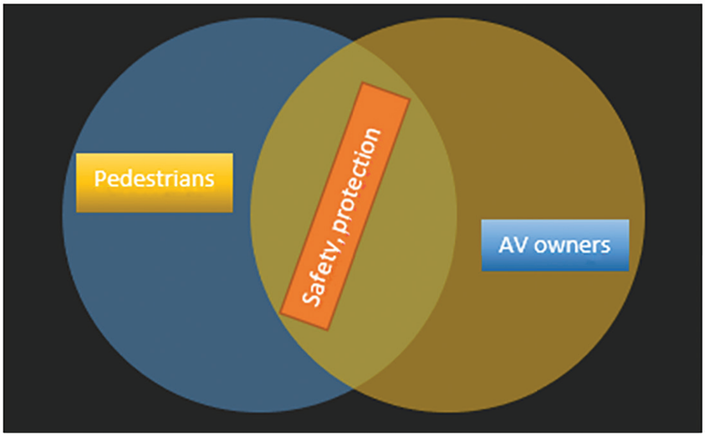


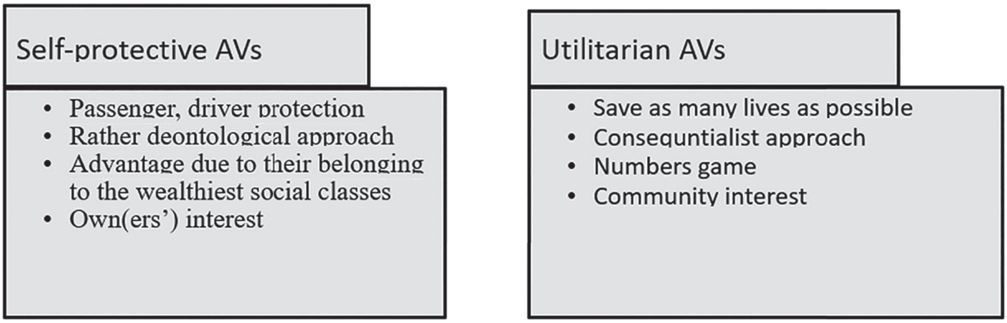
Figure 1. A common cross-section (own creation)

2.2. Self-protective versus utilitarian AVs

We could ask the question why are we talking again about this comparison, self-protective versus utilitarian AVs? The answer is quite simple: because the interests of users and non-users can be argued with these theories.

Self-protective represents a mostly deontological approach that focuses on the passenger's protection. At the base of it is the owner's interest. Otherwise, it is ethically questionable at the same time if we look at the fact that this protective advantage is a privilege of wealthy people since these vehicles cost a fortune. About this we will talk more a bit later in connection with a specific principle.

Utilitarianism approaches the issue from the theory that the highest goal is saving as much life as possible. This depicts the consequentialist aspect and here we are faced with the so-called numbers game. The community interest and the bigger good, less harm are in the center of this and the algorithm operates according to the harm–benefit ratio. The numbers are more important than the use of AVs. At this point it is essential to describe what rule utilitarianism means since this is that type of consequentialist theory that is suitable for this situation. It is a type of utilitarianism that focuses on the rules or principles that govern moral behavior, rather than the outcomes of specific actions. (Miller 2014) According to rule utilitarians, moral rules or principles should be evaluated based on their overall utility or usefulness in promoting the general happiness or well-being of society. The basic idea is that certain moral rules, such as "do not lie" and "do not steal," tend to produce more overall happiness in society than would be produced in a society without such rules. So, it is believed that following these moral rules will lead to the best overall outcome for society as a whole.



Self-protective AVs

- Passenger, driver protection
- Rather deontological approach
- Advantage due to their belonging to the wealthiest social classes
- Own(ers') interest

Utilitarian AVs

- Save as many lives as possible
- Consequntialist approach
- Numbers game
- Community interest

Figure 2. Main differences between Self-protective versus utilitarian AVs

The utilitarian approach can be strengthened by highlighting a few of Patrick Lin's (Lin 2015, 2013a, 2013b, 2013c, 2023) works. He implies that crash optimization, which causes fewer severe injuries, can be applied when programming AVs and this means the as much as it is possible reduction in the severity of consequences. This also follows the consequentialist ethical theory. Crash optimization is driven by a targeting algorithm that calculated the least-victims scenario and with this saves the most lives. But he also says that crash optimization is not enough since accident situations are complex, different and frequently unpredictable such that we need something else, too, not just this algorithm:

Even if consequentialism is the best ethical theory and the car's moral calculations are correct, the problem may not be with the ethics but with a lack of discussion about ethics. Industry, therefore, may do well to have such a discussion and set expectations with the public. Users – and news headlines – may likely be more forgiving if it is explained in advance that self-sacrifice may be a justified feature, not a bug. (Lin 2015, 77)

Lin also draws attention to the different levels of moral obligation that arise from the ownership of AVs, whether they are publicly or privately owned. Due to it, it would be necessary to formulate distinguishable levels of moral – if we are allowed to say so – obligations of the programmed autonomous car.

3. Trust in it or not? Who or what do we trust?

Until we reach at least the 4th level of AVs – which is in progress with promising results – there are more participants in the decision-making and responsibility sides of the AV issues. Of course, we have to deal with this only if something happens, an accident with personal or property damage. Currently, those accountable can be the owners, the producers, and the software developers/engineers (companies). After we cross the line of human-assisted AVs, the 3rd, "conditional automation" level, only the owners would be accountable because any issues will be down to incorrect maintenance, that is, human error is punished. So, improving AVs decreases the number of accountable parties. Theoretically it makes the whole responsibility process simpler, but practically it does not. Why not?

Because programming AVs contains such tasks as are strongly connected with life–death decision-making procedures. How can we implement into a vehicle a moral way of thinking? Are we allowed to choose one quite acceptable scenario for life-saving situations or do we have to accept teaching self-driving cars to do their best according to computing processes? (Service 2021; Gunkel 2018) And here the unsolvable question "whose interest is in the focus?" appears again.

When the 4th level of AVs is achieved, the human factor loses its importance concerning responsibility; humans remain only in the frame of consumers who enjoy this kind of driving service. In this phase of AVs, since machines, algorithms are not able to make decisions, not even moral ones, on their own because of the nature of their operation. They are artificial tools that are programmed and built by humans. (Z. Karvalics 2015) But accidents can happen, and we have to define the responsible ones. This is an extremely complex and difficult task since the implemented programs and algorithms operate in deep learning systems that we are not able to follow in all cases, although we created the programs that enable them to handle all circumstances.

My current topic does not deal with all the decision-making processes but focuses on the moral decisions. Because of the nature of self-driving car programs, we cannot attribute to them the capability of thinking in such a moral way as humans, not even by using a high level of artificial intelligence (AI) because it can only be imitation, following patterns and statistics but not real consideration of dilemmas. So, it can be said in this interpretation that self-driving cars are only tools that operate according to our programmed rules. (Pokol 2017) And the trust we put in these machines is misleading. It should be based on trust toward producers and programmers, but what rules and theories do they follow when creating the systems that are suitable for fulfilling the requirements of safe AVs?

4. An ideal scenario and the principles

4.1. An ideal scenario

Assume that AVs have no bugs, technical malfunctions, or issues with AI programming and deep learning mechanisms, and are functioning perfectly (although this is nearly impossible; we will just imagine that it is so, for the purpose of this thought experiment). In this scenario, we can eliminate the risk of car-related issues. However, accidents may still occur due to human factors, specifically the creators of the AVs. At this point, the principles of consequentialism and deontology come into play and may conflict with each other. In the following analysis, using this ideal mechanical background, I will examine principles that can both strengthen and weaken the introduction of AVs. Although, my goal is not to leave issues unresolved, it is important to recognize that there may be irresolvable contradictions and that an acceptable solution, rather than a perfect one, must be considered when weighing the opportunities and drawbacks.

The principle of justice may be compromised to some extent, but this compromise may not be fully acceptable. This can be compared to the intentions of the principle of beneficence. I will argue that the principle of beneficence is stronger than the principle of justice, but further examination is necessary to understand how this is the case.

4.2. Principle of justice and social inequality

There is no need for special knowledge about the problems of justice and social inequality in the world since we meet them in our everyday lives. This topic is another element on the list that does not decrease current societal issues but, rather, increases them. As I mentioned earlier, since AVs cost a fortune, it is not possible for everybody who has a car to have one. This strengthens inequality and makes the issue of “benevolence” more complex.

The problem with the price, if these can only be bought/owned by wealthier people, is that it won’t be morally managable to make their safety higher in contrast with the safety of poorer people. This privilege is neither acceptable nor fair.

It’s true that more protective, higher-quality tools are more expensive, but in the case of AVs this protection isn’t only passive; it is an active intervention that can also cause harm – not intentionally but as its consequnece – when AVs try to avoid dangerous situations. The question arises, can one buy the security of one's life? The answer is yes, if AVs are available, but from a moral perspective, it is not fair if only the wealthy can afford them. To ensure justice and decrease the previously mentioned factor, AVs should be made accessible to everyone. Furthermore, it would be morally necessary to act like this if the main intention of the society was to decrease accidents, especially lethal ones. It would be an obligation and not a choice. At the same time, it means a kind of force of desirable owners/users to have AVs instead of general cars if the safety and the amount of saved lives are in the center of the regulation.

According to the doctrine of double effect, we can accept the loss because our intention and the consequence differ from each other. But if loss means that we can be victims, the situation cannot be attractive enough. A real moral dilemma appears here, also, that is inevitable: lethal car accidents happen and we cannot cease them completely. It is not possible to act perfectly right because one or other influential factor cannot be changed just handled somehow. Due to the occurrence of safely unsolvable situation, somebody is going to be harmed. We have the ability to decrease the risks of accidents, but we have to define who will be the victims in those cases when somebody has to die. At this point the principle of justice could be involved, but actually it is not capable of helping in those decision-making processes. We could ask: are we allowed to decide who has the right to live and who should die? Not really, but the point is we have to formulate scenarios and rules, taking into consideration the consequences. And if we deal with the inequality factor in the society and the advantage derived from this, we are still in a moral trap.

The message of self-protective AVs seems to be that the lives of the rich are worth more. It is unquestionable and obvious; it is an obvious negative discrimination toward others. The principle of justice is being harmed by the fact of social inequality.

4.3. Principle of beneficence

Beneficence is a tricky expression in the case of real moral dilemmas since it is a principle that cannot be represented perfectly in certain situations. If we introduce utilitarian AVs and oblige the likely wealthy customers to buy this kind of vehicle, it increases their possible harm and does not accurately represent the basic idea of AVs: to save lives. Because of this fact, if the utilitarian result of a probably lethal accident is the death of the AV owner, it obviously cannot be in the interest of the owners.

Although the frequency of deadly accidents is statistically extremely low, sometimes they occur because of the nature of traffic, the behavior of other people on the move, etc. In all deadly cases it is necessary to make decisions and compromises; it cannot be avoided. We have to take into consideration the interests of all the participants. And not just morally but in general, we are not allowed to favor financial interests over human lives.

The point of the principle of beneficence is, hopefully, that it is possible to make a decision that has the best outcome, and it can be a higher number of saved lives.

This shows the theory behind it: saving as many lives as possible. This so-called numbers game has mathematically the most beneficial result, but otherwise we can take into consideration different factors, e.g. quality of life, age, life expectancy, race, and so on. But this does not lead us to equal, fair, and acceptable decision-making processes.

On the one hand, forcing buyers into taking more risk if they want to have an AV is not morally acceptable. But concerning the previously mentioned numbers game, saving as many lives as possible would have to be our duty because of the higher good that can be achieved through that. On the other hand, if we program AVs as self-protective vehicles it is not fair either. Measuring it morally, it is more problematic since we give the privilege to people because they are wealthier, and they can afford to buy these tools.

The consequence of this approach is that utilitarian programmed self-driving cars can discourage buyers from owning one since – even if the chance of serious accidents is extremely low – they could be harmed if they trust in the machine that should serve their comfort, the user's interests and needs, as well as the higher good, they can become victims.

In terms of discrimination, too: if you are rich and want to buy such a vehicle, you have to take the risk that you will be sacrificed for the greater good. It is like a compulsion, or better said, a deal: accepting the risk even if it is extremely low, and enjoying the benefits of AVs hoping nothing serious will happen ever.

It can be seen that the principle of justice and beneficence actually harms from both perspectives but on a different level of severity (in relation to buyers, pedestrians, and others who are participants of traffic).

5. Conclusion – and something else

It is unquestionable that everybody has the right to live, but if a real moral dilemma appears we are obliged to decide how can we influence the consequences and what can be our duties. There is no difference between doing something or not – in the interpretation of non-action is an action also – in connection with making decisions, since both of them have certain results. But in the case of AVs, the consequences can be dramatic. This is why it is necessary to find the best solutions for specific events. *"[S]ome crashes will require AVs to make difficult ethical decisions in cases that involve unavoidable harm"* (Bonnefon 2016, 1573).

It is inevitable that we need to think about moral duty: introducing the available technology is a must if it enables making traffic safer. Not using the technology is equal to omission and unnecessary risk. *"Second—and a more serious problem—our results suggest that such regulation could substantially delay the adoption of AVs, which means that the lives saved by making AVs utilitarian may be outnumbered by the deaths caused by delaying the adoption of AVs altogether"* (Bonnefon 2016, 1575–1576).

Through my writing, I highlighted the difficulties of decision-making and the main obstacles, but I did not recommend specific solutions. Although it would be useful to formulate rules and regulations to handle the morally problematic issues here, I did not do that. That has to be the result of cooperation among certain professional parties, followd by the principles of the main aim: saving lives, making traveling, driving, or just walking safer for everybody. Because of the nature of this kind of decision-making, a less bad solution could turn out to be one of the best ones, since it is not an option to have a perfectly good one.

Accountability and a possible solution?

If we dive into deeper layers and focus on taking risk, it can be said that since introducing AVs is in everybody's interest, it is actually a must. But in this case personal/individual accountability is not an acceptable approach. The consequence of the events is out of our control. Following this way of thinking, it is nonsense to impeach users who do not have any impact on traffic situations. (Bartneck et al. 2021)

Let's think through a logical but theoretical scenario when the interests, needs, and risk-taking factors have been taken into consideration, involving governments as a kind of responsible organization in the following manner. (Tilesch and Hatamleh 2020)

First, it is practical to collect features, needs, and possible consequences of this theoretical case. The collective interest of introducing AVs is the core of the whole situation.[2] The reason is not new, because of the life-saving benefits of using more-protective tools in everyday life. Maximizing the effectiveness of this traffic opportunity – in the sense of the numbers game – governments should support development processes and make AVs available not just to wealthy people. At this point it is necessary to highlight accountability, which is one of the most essential questions. If we want to introduce AVs because of safety reasons and we do not want to punish users and non-users with the unlucky, rare, and undesired, unwanted consequence of an unfortunate accident, it is more realistic to transfer the financial and other consequences to the organization that "forced," advertised, and supported the use of these tools.[3] The intention of representing AVs is clear: the safety risks are small compared to general cars and drivers and that can drive a theoretically acceptable risk-taking level.[4]

Concerning other effects on the government side, it requires an increasing amount of financial resource investment if equal availability is the focus (Héder 2020). At this point we can formulate another question: should everybody get such a car (AV) who has a general one or who wants to have a car in general? Being prepared for this financially and measuring legitimacy demands are complex issues.

[2] And it was through the whole essay also, so please forgive me for the repetition, but it is not acceptable to not mention it again right here.

[3] Although it is in the society's interest.

[4] If it will ever have such a level. Since we are talking about human lives, it is not a perfect formulation of the thoughts, but in a comparative sense the fewer victims, the more acceptable and supportable it is. Without comparison, no victim is acceptable ever. But this statement cannot be represented in practice with the current conditions.

Instead of this economic calculation I would recommend a golden mean that could act as a kind of solution for utilitarian versus self-protective self-driving cars: the Autonomous Vehicle-Car Sharing System (in the following: AV-CSS).

In this idea, governments have the responsibility to operate and maintenance AV car fleet This makes it clear who will be responsible and accountable in accident situations. If somebody wants to use an AV, they have the chance, since it should be affordable, but the user takes the risk that they will become an innocent victim in a lethal accident should the AV try to act according to consequentialist rules.[5] On account of the nature of "AV safety" probability, the risk taken is significantly lower than in general cases. The more AVs are participants of public transportation and traffic, the less probability of lethal accidents there is. Because of the kind of common network in relation to knowledge base that teaches AVs about each traffic situation, it makes this kind of transport safer, and it can become more calculated with less risk, so there is less likelihood of harm and in the ideal case, finally, it should be possible to eliminate lethal accidents entirely.

With this AV-CSS, forcing anybody to use an AV is avoidable, but it makes it possible for everybody to live with the opportunity and enjoy the comfort of these tools at the same time. Individual free decision-making opportunity is preserved and represented.

Moreover, car sharing has other beneficial consequences, especially if AVs are electrical cars. Less environmental load, fewer emissions, less expense, fewer cars owned. All of these mentioned results of AV-CSS have a huge impact on society and on the environment as a whole. So, from my point of view, making steps toward this kind of less consumer-centric behavior is worth considering. It is in the interest of all of us. Not by the way, it can solve the analyzed issue of whether self-protective or utilitarian AVs should be introduced? This idea gives a potential way for handling the issue and giving the opportunity to decide freely: to use or not to use. The possibility of choosing the lesser evil should be open to individual decision. What we need is transparency, human oversight, and public and governmental support. Taking into account all of these factors will not solve all of the issues, but with help they can be decreased and narrowed down.

[5] As has become obvious, this scenario operates according to consequentialist theory – saving as many lives as possible. The reason is that any of the approaches possibly have bad consequences, but in this case the user decides whether they take the relatively low risk and use AVs or not.

References

Bartneck, Christoph, Christoph Lütge, Alan Wagner, and Sean Welsh. *An Introduction to Ethics in Roborics and AI*. Cham: Springer, 2019. https://library.oapen.org/viewer/web/viewer.html?file=/bitstream/handle/20.500.12657/41303/2021_Book_AnIntroductionToEthicsInRoboti.pdf?sequence=1&isAllowed=y

Bonnefon, Jean-Francois, Azim Shariff, and Iyad Rahwan. "The social dilemma of autonomous vehicles." *Science* 352, no. 6293 (June 24, 2016): 1573–1576. https://www.DOI:10.1126/science.aaf2654

Braun, Joachim von, Margaret S. Archer, Gregory M. Reichberg, and Marcelo Sánchez Sorondo. *AI and Humanity*. Cham: Springer, 2020. https://doi.org/10.1007/978-3-030-54173-6

Britton, Jill. "What are the autonomy levels for autonomous vehicles?" Accessed July 4, 2022. https://www.perforce.com/blog/qac/6-levels-of-autonomous-driving?utm_source=googleadwords&utm_medium=cpc&utm_campaign=QAC-Dynamic&utm_adgroup=Dynamic&gclid=Cj0KCQjwn4qWBhCvARIsAFNAMiidztbql699pW39khOyco-3i3mBYr-4-v3lO2G_OfOdhgP0e9Kez6UaAg19EALw_wcB.

Héder, Mihály. "A criticism of AI ethics guidelines." *Információs Társadalom* XX, no. 4 (2020): 57–73. https://dx.doi.org/10.22503/inftars.XX.2020.4.5

Gunkel, David J. *Robot Rights*. Cambridge, MA: MIT Press, 2018.

Lin, Patrick. "Why ethics matters for autonomous cars." In *Autonomes Fahren*, edited by Markus Maurer, J. Christian Gerdes, Barbara Lenz, and Hermann Winner, 69–85. Heidelberg: Springer, 2015.

Lin, Patrick. "The ethics of saving lives with autonomous cars is far murkier than you think." *Wired* (July 30, 2013a). Accessed Jan. 4, 2023. http://www.wired.com/2013/07/the-surprising-ethics-of-robot-cars/

Lin, Patrick. "The ethics of autonomous cars." *The Atlantic* (October 8, 2013b). Accessed Jan. 4, 2023. http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/

Lin, Patrick. "Why the drone wars matter for automated cars." Accessed Jan. 4, 2023. http://stanford.io/1jeIQuw.

Maurer, Markus J., Christian Gerdes, Barbara Lenz, and Hermann Winner. *Autonomous Driving*. Heidelberg: Springer Nature, 2016.

Miller, Dale E. "Rule utilitarianism." In *The Canbridge Companion to Utilitarianism*, edited by Ben Eggleston, and Dale E. Miller, 146–165. Cambridge: Cambridge University Press, 2014.

Pokol, Béla. "A mesterséges intelligencia." *Információs Társadalom* XVII, no. 4 (2017): 39–53. https://dx.doi.org/10.22503/inftars.XVII.2017.4.3

Service, Robert F. "Learning curve, new brain-inspired chips could provide the smarts for autonomous robots and self-driving cars." *Science* 374, no. 6563 (October 1, 2021): 24–25. https://www.science.org/content/article/new-brain-inspired-chips-could-soon-help-power-autonomous-robots-and-self-driving-cars

Tilesch, George, and Omar Hatamleh. *Between Brains: Taking Back Our Future in the AI Age*. PublishDrive, United States, 2020.

Wallach, Wendell, and Colin Allen. *Moral Machines: Teaching Robots Right and Wrong*. Oxford: Oxford University Press, 2009.

Z. Karvalics, László. “Mesterséges intelligencia – a diskurzusok újratervezésének kora.” *Információs Társadalom* XV, no. 4 (2015): 7–41. https://dx.doi.org/10.22503/inftars.XV.2015.4.1