Choose act or rule utilitarianism to motivate collision-avoidance protocols for autonomous vehicles? An Analytical Review.

Ziyi Liang

University of New South Wales

Nowadays, autonomous vehicles (AVs) are not yet a mature technology, and there is a considerable controversy; no one knows if they should leave the safety of human lives in the hands of machines. Nevertheless, the market potential for fully AVs is quite prominent in the coming decades. Fully AVs will bring disruptive improvements in areas such as enhancing highway safety, easing traffic congestion, and reducing air pollution. AVs are not a single new technology but an integration of various technologies. It requires not only hardware development but, more importantly, how engineers should design the ethical standards and rules of human society into the algorithm. That is, how AVs should make reasonable judgments in times of unexpected situations. The "ethical factor" is crucial in allocating the risk among drivers, pedestrians, and public property. For engineers and the general public, the AVs' decision-making system must consider human ethics in every decision. When faced with ambiguous ethical dimensions, we generally focus on complying with the law while minimizing the damage caused by specific consequences. This strategy is more popular because it is a decision based on legal regulations, and developers have little effort to explain the behaviour of the AVs. In this way, the responsibility for defining ethical behaviour is shifted to lawmakers. However, the law generally does not give specific instructions on "what to do in an emergency," especially if the only two options available are both life-threatening. In such a particular case, it is left to the machines to decide which option will produce the best result. Therefore, this essay will discuss how choosing a different utilitarianism to motivate collision-avoidance protocols for AVs can lead to different protocols and why both utilitarian incentive protocols seem flawed.

The details of act and rule utilitarianism and their differences will be introduced first, then compare the consequences of using them to guide collision avoidance protocols, and finally, reasons will be given as to why neither utilitarianism is well suited.

Bentham (1789) believed that happiness or pleasure is homogeneous and, therefore, can be compared and summed up between people. Hence, the basic idea of utilitarianism is that ethical behaviour or institutions should promote "the greatest happiness of the greatest number that is the measure of right and wrong". One of its more prominent features is that it is permissible and advocated to sacrifice individual happiness and even life to increase collective happiness. The trolley problem, for example, is a classic example of a moral dilemma, and the sacrifice of one person to save five is a move promoted by utilitarianism. To summarize, utilitarianism is based on maximizing benefits, and whose interests will be harmed in the process is not a factor in its consideration. In Bentham's own words, the utilitarian principle is to express approval or disapproval of any action according to the tendency of the matter to increase or decrease the pleasure of the stakeholders, that is, to see whether the action will bring them pleasure or pain (Bentham & Bowring 1843). Another case in point is that many science fiction movies define the reason for robots to wipe out humans is that the robots have discovered through their calculations that the most direct way to best save the planet is to wipe out humans. This is a vivid example of utilitarianism.

Different branches of utilitarianism have been developed later, such as act utilitarianism.

Unlike other schools of utilitarianism, act utilitarianism emphasizes the question, "what can be done to promote the value of happiness for all people in this situation at this time? "Rather than asking what effect the extension of this moral law to everyone would have on the happiness value of the whole. A simple example is that lying is considered unethical in everyday life. However, in certain situations, the act utilitarian would consider lying to be ethical behaviour. An illustration would be an agent lying to keep state secrets. This is denying the significance of moral rules, arguing that all people and their situations are different and that it is impossible to set uniform moral rules for behaviour: in choosing behaviour, people must gauge their own situation and act directly on the utilitarian principle, i.e., choose an outcome that is good not only for themselves but also for the greatest number of all those associated with it.

In contrast, there is another branch of "rule utilitarianism". Rule utilitarianism believes that the greatest pleasure value can be generated if everyone always follows a set of moral rules, such as the rules of the road. If everyone obeys the rules of the road, everyone can drive happily and safely. Emphasis on the morality of an action is when it is beneficial to the majority of people and the simultaneous adherence to a particular rule or code of conduct that leads to the greatest good or happiness of the people.

A significant difference between act and rule utilitarianism is that in act utilitarianism, the consequences are in the "act," whereas in rule utilitarianism, the consequences are in the "rule". Act utilitarianism evaluates current behaviour through the actual

consequences of current behaviour. In contrast, rule utilitarianism evaluates current behaviour through the consequences of the general practice of behaviour (through all people, both in the past and future). For example, if cheating on an exam gets you a high score and makes your parents happy, the Act Utilitarian will cheat because it maximizes the benefits for everyone. The rule utilitarian would not do so because if everyone cheated, no one would be qualified for college, and grades would be meaningless. This would lead to a reduction in the total value of benefits. A rule utilitarian follows a rule and can make the decision to maximize the gain by following that rule. In this case, it is the rule that no one can cheat, which may not result in a maximum gain in the short term, but is the best decision in the long term. Note that the rule utilitarianism will focus on long-term benefits. It is not good enough to follow this rule if it brings maximum benefits in the present but no guaranteed benefits in the future. So the rule utilitarian can, in some cases, not maximize the present gain and thus ensure that the long-term gain is stable and does not decrease.

There is a classic category of ethical dilemmas that AVs may encounter. That is, if an AV is driving on the road, but there are suddenly three children crossing the road in front of it, and there are heavy stone piers to the left and right of the vehicle. The AVs must now choose whether to crash into the children in front or ignore the owner's life and turn sharply into the stone pier. Thus sacrificing the only occupant to avoid the death of the three children. At this point, the act utilitarian motivated protocol would consider the probability of surviving your crash into a stone pier and the consequences

of three children dying. A trade-off will be made to sacrifice the owner of the car. This is because the act utilitarian choice results in the greatest good for the greatest number of people in the situation and the greatest preservation of life after considering the current situation and the consequences of the act.

On the other hand, the rule-utilitarian protocol additionally considers whether the general overall consequences of the choice to sacrifice the owner of the vehicle would be of the greatest benefit if the current situation were to occur many times in the future. That is, what would be the consequences to society and all humans if the choice to sacrifice the owner in this situation were to become a universal behaviour that every AV would do. In this instance, if the choice is made to sacrifice the owner, then there could be a blow to the AVs industry and the consumer's desire to buy would be diminished. This is because a segment of the population will not accept this choice of a machine choosing to sacrifice itself in a crisis situation without the user's consent. According to the World Health Organization (2018), road traffic injuries caused an estimated 1.35 million deaths worldwide in 2016. While it is not currently proven that AVs are safer than human drivers, the continued development of AVs may eventually lead to a significant reduction in crash rates. However, if AVs fail to become widespread, it seems that in the long run, they will also lead to more deaths and a decrease in the general well-being of society. Any choice that hinders the development of AVs indirectly harms those who may be killed by regular vehicles in the future. In the eyes of rule utilitarianism, the overall benefit to the future of humanity outweighs any risk

posed by current unethical vehicles. Rules-utilitarian protocols need to ensure that the behaviour of AVs is acceptable to vehicle owners, which is also essential to gaining public trust in these new technologies.

In collisions involving moral dilemmas, purely Act Utilitarian protocols simply minimize social loss, while purely Rule Utilitarian protocols may be immoral in individual cases, neither of which seems to be the best option. The rule utilitarianism-inspired collision avoidance protocol mentioned above argues that occasional immoral AV can be ignored for the overall benefit of the future of humanity, for which Goodall (2014) criticizes: if the happiness of one group may be at the cost of another, then it is socially unacceptable even if it is an improvement for humanity as a whole. Furthermore, in an experiment by Mayer et al. (2021), they described the classical moral dilemma of AVs to participants, who had to choose whether it should be "at the expense of the pedestrian" or "at the expense of the passenger". The result was that 60 per cent of the participants preferred to save the maximum number of lives, whereas also 40 per cent preferred the option of self-preservation, even if it meant sacrificing pedestrians (Mayer et al. 2021). They also mentioned that perspective strongly determined participants' choices, with those prompted for the passenger perspective consistently indicating a reluctance to sacrifice passengers, while pedestrians indicated a willingness to sacrifice passengers. This is yet another indication that neither purely utilitarian protocols of act or rules may be acceptable to the general public's moral values.

Thinking about it from another perspective, we may not need to cede command; the moral choice of AVs should be chosen in advance by the people riding the vehicle. That is, a moral bias setting should be provided for AVs. This setting would allow passengers to choose their own moral preferences in an emergency within the bounds of legality. For example, the choice would be more toward act utilitarianism or rule utilitarianism or egoism. Thus the AV's decision in the face of an ethical dilemma would depend on the user's customized choice, and the responsibility would be assigned to the user rather than to the engineer and the manufacturer. According to the suggestion made by Contissa et al. (2017), if users set the moral bias of AVs to extreme egoistic modes, then those who preselect this preference should pay higher premiums. Of course the extreme egoistic behaviour by AVs here should also be an AV behaviour that is allowed under the legal framework and does not expose the user to criminal or civil liability. An collision-avoidance protocol that can set moral bias can better advance the AV industry. Legislators could set boundaries for a moral dilemma, allowing users to choose their preferences within the range.

To summarize, in a moral dilemma, AVs must choose the lesser of two evils, which raises many ethical and legal questions. In this essay, two different utilitarian motivations for collision avoidance protocols are explored in terms of what choices and behaviors they would make. These are act utilitarianism, which makes the behavior that yields the greatest benefit to the well-being of all in the current situation, and rule

utilitarianism, which emphasizes following the rules and thinking about the bigger picture. Both seem to make sense in their respective contexts but are not accepted by all. Leaving the moral choice to the user seems to be a solution that can neutralize the disadvantages of both strategies. At least AVs with such a protocol would be more prevalent in the market and good for the development of the AV industry as a whole. Thus, neither act nor rule utilitarian motivated AV collision avoidance protocols seem to be the best choice.

References

Bentham J. & Bowring J. (1843). The works of jeremy bentham. W. Tait; *Simpkin Marshall.*

Bentham, J. (1789). An Introduction to the Principles of Morals and Legislation. *Clarendon Press,* Oxford. http://dx.doi.org/10.1093/oseo/instance.00077240

Contissa, G., Lagioia, F., & Sartor, G. (2017). The Ethical Knob: ethically-customizable automated vehicles and the law. *Artificial Intelligence and Law,* 25(3), 365-378.

Goodall, N. J. (2014). Machine ethics and automated vehicles. *In Road vehicle automation* (pp. 93-102). Springer, Cham.

Mayer MM, Bell R, Buchner A (2021) Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles. *PLoS ONE* 16(12): e0261673. https://doi.org/10.1371/journal. pone.0261673

World Health Organization. (2018). Global status report on road safety 2018: summary. *World Health Organization*, Geneva.