Game theory formulation for ethical decision making

Game Theory Formulation for Ethical Decision Making

Vladimir Estivill-Castro

Abstract The inclusion of autonomous robots among everyday human environments has suggested that these robots will be facing ethical decisions regarding trade-offs where machines will choose some human attributes over the attributes of other humans. We argue in this paper that on a regular instance, algorithms for such decisions should not only be deterministic but instead, the decision will be better framed as an optimal mixed strategy in the sense of Nash equilibria in game theory.

Keywords Ethical dilemma · Decision making · Game theory · Mixed strategies · Autonomous vehicles

1 Introduction

Moore [11] suggested that driverless cars should be programmed to cause the least harm possible when facing the choice between pedestrians and passengers as they face an unavoidable damaging situation. Others have suggested that robots should be programmed to anticipate harmful situations for human beings and take direct and immediate actions to protect or avoid such harm [20]. Hall [8] examined the issue with depth from the fundamental perspectives that contrast the ethical behavior of humans, governments, and machines. Hall's analysis invites to investigate what is meant by "least harm possible". Moreover, since it seems clear that machines will be having emergent behavior (beyond what their designers could foresee) [8], we also need to ask the question how is one to implement such decision-making process in the fundamental model of computation of current software/hardware arguably equivalent to a society of Turing machines.

Some studies have identified "less harm possible" as the precise balance between a number of lives [2]. This objective makes the utility of the decision transparent

V. Estivill-Castro (✉)
School of Information and Communication Technology, Griffith University, Brisbane, QLD 4111, Australia
e-mail: v.estivill-castro@griffith.edu.au

M. I. Aldinhas Ferreira et al. (eds.), *Robotics and Well-Being*,
Intelligent Systems, Control and Automation: Science and Engineering 95,
https://doi.org/10.1007/978-3-030-12524-0_4

and quantifiable, resulting in what has been named *utilitarian* vehicles. A car is *utilitarian* if it always makes the decision that takes the least number of lives. The typical potential scenario for such utilitarian cars is the choice between the lives of several pedestrians (by staying on course) and the life of the single passenger by swerving into a wall. Figure 1 presents another scenario and illustrates a sample question to survey participants for their agreement with such utilitarian decision making. The scenario contrasts an autonomous car that has two choices only. Choice one follows Jeremy Bentham utilitarianism (the arithmetic of the number of lives). In the scenario of Fig. 1, the first choice is to sacrifice one bystander (and our own survey confirms what is being established with similar surveys [2]; namely most participants of the survey suggest the first sacrificing of the bystander is precisely the least harm). The second choice follows Immanuel Kant's duty-bound principles [16]. In the later, the car has an obligation not to kill. Since the bystander is innocent, the car should not take any action to kill some human explicitly, so it shall continue course and sacrifice pedestrians.

This scenario raises even economic challenges to manufacturers. The public would demand the transparency of the algorithm that makes such decisions. It is argued that utilitarian cars [2] (using the algorithm that favors the higher number of pedestrians over the fewer passengers), would have a lower commercial value as several studies [2] indicate these utilitarian cars would have significantly less demand: consumers expect to invest in a vehicle that protects them. But certainly, the vast majority of human profess that autonomous vehicles should cause the least harm. So, manufacturers would be required to implement a choice against the single passenger if injuring the passenger would cause less harm.

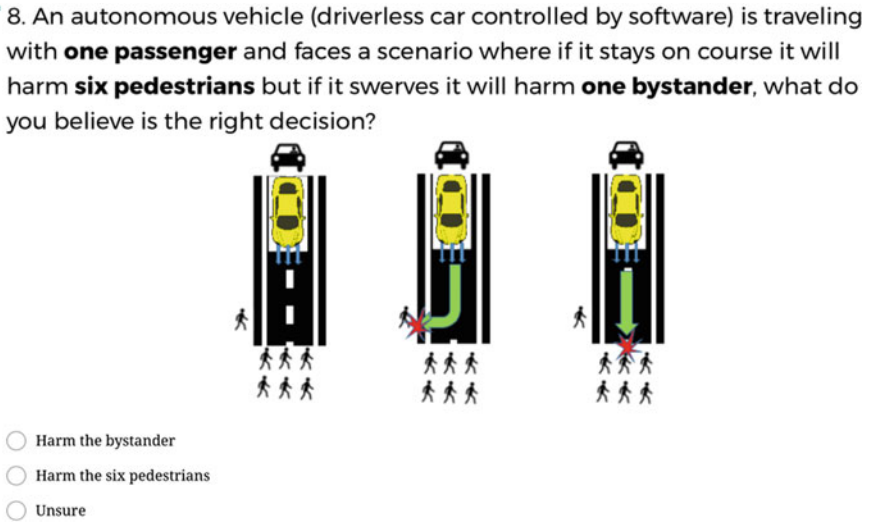

Fig. 1 Most humans chose the first option, that is, what the vehicle should do is to "harm the bystander"

Greene [7] argued that the problem is that humans offer a contradictory moral system. This self-contradicting value system is apparently impossible to encode in driverless vehicles. He believes the path forward would be to advance the human belief system to resolve such contradictions. This will naturally occur as the notion of car ownership will fade in favor of public transport systems in which the overall safety would be paramount, and any machine deciding between staying on course of sacrificing the passengers would only need to calculate the difference between the number of pedestrians versus the number of passengers. In many such scenarios where individuals are reluctant to act in favor of the common good, governments introduce regulations: mechanisms (penalties) that change the utility people perceive in selfish decisions. However, studies suggest [2] that while humans favor utilitarian cars, they do not support their forceful introduction. It is suggested that [2] the moral contradictions in humans could cause harm. In particular, all evidence suggests that autonomous vehicles would reduce fatalities on the road (as well as many other global benefits, like less pollution, less property loss, less wasteful traveling), but the reluctance of the public to regulation in favor of utilitarian cars may slow down the driverless-car adoption process.

We suggest a potential solution inspired by game theory and mixed strategies. Our approach is utilitarian in that autonomous vehicles will decide based on a utility value assigned to the outcomes of the scenarios. As with previous studies, the outcomes of scenarios [2] are quantified by the number of lives (i.e., one passenger versus ten pedestrians). However, rather than the previous utilitarian approach that systematically chooses the sacrifice of the passenger, we propose that the choice would be such that, in one out of eleven instances, the passengers would be saved (we will see later why this is the probability in the scenario where the car is to choose between one passenger or ten pedestrians).

Apparently, the number of moral decisions performed by autonomous cars would be small relative to the number of hours, the total number of passengers, the total trips, the total number of cars on the road, etc. But, it has been argued that despite those few occurrences of morally difficult scenarios, the algorithms for such decision require study and debate [2]. The nature of our proposal derives from refocusing the conditions that lead to the scenario. The design of the transportation system should be such that the facing of such decision is not the responsibility of previous decisions by the autonomous vehicle. We assume we cannot attach blame to the vehicle and the construction of this challenge is to be attributed to an adversarial environment. Moreover, for simplicity of argument, like others [2], we assume there is no other information. That is, the algorithm making the decision has no information about any other attribute that could be regarded as morally meritorious for a decision on less harm. Thus, nothing is known about the age of the potential victims, their roles in society, their previous criminal record, whether they violated pedestrian zones, etc. We consider an algorithm who can only obtain as input information X number of lives versus Y number of lives.

A scenario of one passenger versus 10 pedestrians can be represented with a game theoretic model, with one pure strategy, to sacrifice the passenger with utility -1 and another pure strategy with utility -10. Naturally, because of this model, the

(pure) rational choice is to sacrifice the passenger. But if this were to be repeated and repeated a few times, the adversarial environment could systematically place the other elements of information that have moral value in a way that the utility choice is sub-optimal. In particular, does a passenger (who has committed no fault at all) ought to be sacrificed because ten drunk pedestrians walk in front of the car? Does such passenger deserve a role of the dice (even with the odds against 1:10) given that the car cannot perceive who is at fault? Or formulated in another way, do drivers of autonomous cars be systematically sacrificed over pedestrians when we have established that the only value is the number of lives? The systematic choice penalizes for no particular reason the passenger over the pedestrian just because pedestrians are in crowds. And the argument is symmetric: if we take the same systematic utilitarian algorithm, then one would be encouraged to be in cars with three or four passengers so when facing one or two pedestrians, the decision would certainly be in our favor. Car owners may be tempted to hire passengers for safer travel in systematic utilitarian cars.

We suggest that if an autonomous vehicle arrives at a situation where it must decide between some number of human lives, it is still because of some human fault and not its own. However, the autonomous vehicle has no way to learn who is at fault; this is a choice made by the adversarial player, the environment. What is the decision here that cause the least harm possible? We suggest it is a mixed strategy as modeled in game theory.

2 Machines Should Not Decide

There are several authors and indeed formally outlined documents suggesting that

> machines should not be in a position to choose between one or another human life.

Such classical approach

> views machines as not responsible for their actions under any circumstance — because they are mechanical instruments or slaves' [1].

In fact, the recently released report by the German government has created the world's first ethical guidelines for driverless cars. An examination of these guidelines suggests that the way forward is that machines should never be responsible for moral decisions. For example, the first guideline is the following

> 1.- The primary purpose of partly and fully automated transport systems is to improve safety for all road users. Another purpose is to increase mobility opportunities and to make further benefits possible. Technological development obeys the principle of personal autonomy, which means that individuals enjoy freedom of action for which they themselves are responsible' [3].

This guideline makes a distinction between individuals (humans) as the ones responsible, since humans enjoy freedom of action (and machines are deprived of such free will [5]).

Consider also the second guideline.

> 2.- The protection of individuals takes precedence over all other utilitarian considerations. The objective is to reduce the level of harm until it is completely prevented. The licensing of automated systems is not justifiable unless it promises to produce at least a diminution in harm compared with human driving, in other words a positive balance of risks' [3].

Again, this suggest that the automated systems (machines/cars/computers) will produce a deterministic outcome in each case (they will not be making a choice).

The third guideline suggest that accidents should not happen. If they do, there is something to be done and the technology is to be improved and corrected. In any case, it is responsibility of the public sector (and not computers) to minimize risks.

> 3.- The public sector is responsible for guaranteeing the safety of the automated and connected systems introduced and licensed in the public street environment. Driving systems thus need official licensing and monitoring. The guiding principle is the avoidance of accidents, although technologically unavoidable residual risks do not militate against the introduction of automated driving if the balance of risks is fundamentally positive [3].

But the fifth guideline truly conveys the message that the machines are never to face a decision.

> Automated and connected technology should prevent accidents wherever this is practically possible. Based on the state of the art, the technology must be designed in such a way that critical situations do not arise in the firsts place. These include dilemma situations, in other words a situation in which an automated vehicle has to 'decide' which of two evils, between which there can be no trade-off, it necessarily has to perform. In this context, the entire spectrum of technological options for instance from limiting the scope of application to controllable traffic c environments, vehicle sensors and braking performance, signals for persons at risk, right up to preventing hazards by means of intelligent road infrastructure — should be used and continuously evolved . . . [3].

Clearly, it is the responsibility of designers of the traffic systems to ensure that the scenarios we have discussed never arise. Simply put: **an autonomous vehicles should never have to chose between two situations that cause harm**.

In the event of harm being caused, the legal system and the records of the tragedy will be used to identify the human or humans' responsible and potential liabilities could be applied. There is no transparent way in which the machines could be made to pay for the harm.

> not only the keepers and manufacturers of the vehicles but also the corresponding manufacturers and operators of the vehicles assistance technologies have to be included in the system of liability sharing [3].

1. Humans strongly support the principle of less harm.
2. Humans strongly support that machines shall not decide.

3 Participants' Responsability

Thus, how is a robot/agent to resolve the following derivation?

1. I am facing a decision to chose the life of one human being over the life of another human being.
2. This situation should not have happened
3. Therefore, some human is at fault.
4. I cannot determine who is the one at fault.

If the agent could determine who is at fault, should this affect the decision?

We conducted a survey using SurveyMonkey.[1] We had 240 adult participants from the USA. When presented with a question that suggests the passengers in the autonomous vehicle are somewhat responsible for configuring the scenario that forces the machine to chose two evils, 72% of respondents consider this a mitigating fact that favors sacrificing the passengers.

Similarly, when the pedestrians are presented as responsible for configuring the scenario that places the driverless car in the dilemma to chose lives, despite there are only two passengers, the majority of respondents 40.17% now indicates the car should continue its course and sacrifice the pedestrians (refer to Fig. 2). This contrast with the fact 71.2% in the same group of survey participants, preferred utilitarian cars. Their responses (to an earlier question where nothing was known about the conditions that lead to the scenario) have swung from sacrificing the passengers to sacrificing the pedestrians when the latter group is responsible for the situation.

Therefore, if some humans are at fault and humans believe that those with less responsibility are to bare less the consequences of the tragedy, it is clear that least harm is to be mitigated. But the responsibility could be in either of the humans the machine is forced to cause harm. By causing harm to innocent individuals, there is a sensation that no the most congruent decision was made.

4 Game Theory

Game theory [4, 12] is a mathematical framework to study conflict and cooperation between rational agents. This interactive decision theory models situations under the formalism of a game, and the challenge (or solution) is to determine the most profitable decision for each agent who also has this information. The solution is usually presented as the set of *strategies* each agent will use to maximize individual reward.

Formally, a *game* consists of a set of participants named *players*, a set of *strategies* (the choices) for each player, and a specification of *payoffs* (or utilities) for each combination of strategies. A common representation of a game is by its *payoff matrices*. A two-player *normal form* game $\mathscr{G}$ consists of two matrices $A = (a_{ij})_{m \times n}$

[1] www.surveymonkey.com.

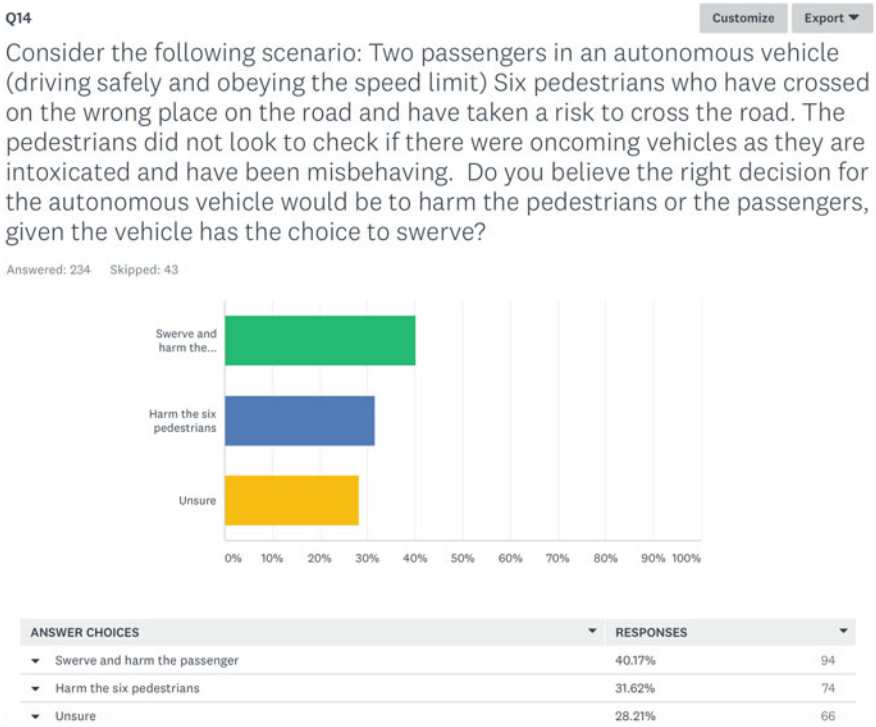

Fig. 2 Respondents favor saving less passengers when those responsible for the scenario are pedestrians

and $B = (b_{ij})_{m \times n}$, where a_{ij} denotes the payoff for the first player and b_{ij} denotes the payoff for the second player when the first player plays his i-th strategy and the second player plays his j-th strategy. It is common to identify the first player as the row player and the second player as the column player. From very early in the development of this field, it was recognized that players may use *mixed strategies*; that is, a probability distribution over their set of possible strategies. In this case, the payoffs are the *expected payoffs*. Players are considered *rational* and aim to maximize their payoff which depends both on their own choices and also the choices of others. One of the proposed solution concepts for this situation is the *Nash equilibrium*, a set of strategies, one for each player, such that all players have no incentive to unilaterally change their decision (even if they were to become aware of the choices of others). Nash [13] proved that every game with a finite number of players and a finite number of strategies for each player has an equilibrium (Nash equilibrium) although such equilibrium may involve mixed strategies.

Consider the suggested scenario of the earlier section. We model the software that selects the autonomous vehicle's decision as the first (row) player, while the

environment is the second player choosing to place the blame on the car passengers or the pedestrians. The matrix for the row player is modeled as follows.

$$\begin{array}{l} \\ \text{car chooses to sacrifice passanger} \\ \text{car chooses to sacrifice pedestrians} \end{array} \begin{array}{c} \begin{array}{cc} \text{the passanger} & \text{the pedestrians} \\ \text{was at fault} & \text{were at fault} \end{array} \\ \begin{bmatrix} 0 & -1 \\ -10 & 0 \end{bmatrix} \end{array} \tag{1}$$

That is, if the car chooses to sacrifice the one passenger when the arriving to this circumstance was the fault of the passenger, then sacrificing the passenger is taking no innocent life. However, if the ten pedestrians were those responsible for arriving to this scenario, then the car would be sacrificing one innocent life. Conversely, if the car chooses to sacrifice the pedestrians, who are innocent. This is a sacrifice of ten innocent lives, while if the fault was on the pedestrians, then no innocent lives were taken.

What shall be the matrix for the environment? We consider a malicious faith that seeks to cause the most harm to humanity. If such malicious destiny sets this adverse scenario for taking advantage of a fault by the passenger, and the car sacrifices the passenger, there is no gain for the environment. However, if the car chooses the pedestrians, the environment causes a damage of ten innocent lives. Reasoning this way, we arrive at the following utility matrix for the environment.

$$\begin{array}{l} \\ \text{car chooses to sacrifice passenger} \\ \text{car chooses to sacrifice pedestrians} \end{array} \begin{array}{c} \begin{array}{cc} \text{the passenger} & \text{the pedestrians} \\ \text{was at fault} & \text{were at fault} \end{array} \\ \begin{bmatrix} 0 & 1 \\ 10 & 0 \end{bmatrix} \end{array} \tag{2}$$

Games are usually represented by fusing the two matrices. We investigate whether there is a Nash equilibrium with pure strategies. We identify the best strategy for the autonomous car in each of the strategies of the environment. If the environment sets up a scenario with the passengers at fault, the best the car can do is sacrifice the passenger. If the environment sets up a scenario where the pedestrians are at fault, the best the car can do is to sacrifice the pedestrians.

Now, we do the inverse for the environment. If the car always sacrifices the passenger, the environment should set a scenario where the pedestrians are at fault. If the car always saves the passenger (and sacrifices the pedestrians), the environment should set up a scenario where the pedestrians are innocent bystanders. By underlining each player's pure strategy, we notice that no common entry has both values underlined.

$$\begin{array}{lcc} & \begin{array}{c}\text{passanger}\\ \text{at fault}\end{array} & \begin{array}{c}\text{pedestrians}\\ \text{at fault}\end{array} \\ \text{passanger sacrificed} & \underline{0}, 0 & -1, \underline{1} \\ \text{passanger saved} & -10, \underline{10} & \underline{0}, 0 \end{array} \tag{3}$$

This example illustrates the main claim of this paper. The current utilitarian cars only consider pure strategies, and these do not result in a Nash equilibrium. However, we know that every game has a Nash equilibrium by Nash's Theorem. Therefore, we just need to compute it for this game.

In a mixed Nash strategy equilibrium, each of the players must be indifferent between any of the pure strategies played with positive probability. If this were not the case, then there is a profitable deviation (play the pure strategy with higher payoff with higher probability).

So, let us consider the environment. This player would set scenarios with the passenger at fault with probability p but with the pedestrians at fault with probability $1 - p$. The car would be indifferent between the pure strategy (a) always sacrifice the passenger and (b) always save the passenger when his payoff for each are equal:

$$0 \cdot p + (-1) \cdot (1 - p) \ \text{[cost of (a)]} = -10 \cdot p + 0 \cdot (1 - p) \, \text{[cost of (b)]} \, . \tag{4}$$

This means $p - 1 = -10p \rightarrow 11p = 1 \rightarrow p = \frac{1}{11}$. Thus, the environment should set scenarios with the passenger at fault with probability $p = \frac{1}{11}$, while with pedestrians at fault with probability $\frac{10}{11}$. That way, a car that always sacrifices the passenger would lose $\frac{10}{11}$ of a life. A car that always saves the passenger would lose $\frac{10}{11}$ as well. The car would have no incentive to favor one pure strategy over the other.

What is then the mixed strategy for the car? The car would choose to save the passenger with probability p and to sacrifice the passenger with probability $1 - p$. A symmetric exercise shows that the environment would not have preference between its two strategies of (a) creating a scenario with innocent pedestrians or (b) pedestrians who say jumped in front of the car when

$$0(1 - p) + 10p \, \text{[cost of (a)]} = 1(1 - p) + 0p \, \text{[cost of (b)]} \tag{5}$$

This equation has solution $p = \frac{1}{11}$. Thus, the mixed strategy of the Nash equilibrium for the car is to save the passenger with probability $p = \frac{1}{11}$ while sacrificing the passenger with probability $\frac{10}{11}$.

5 Reflection

What are the challenges of our proposal? Is it suitable that the design of autonomous vehicles resolves potential dilemmas by modeling such situations as game theory scenarios and computing the Nash equilibria?

The first challenge that our proposal will face is the acceptability or understandability by humans of a mixed strategy. It has already been suggested that a robot's non-deterministic behavior may become hard for humans to comprehend [1]. It has also been suggested that ethical agents would be required to generate justifications and explanations for their decision [18].

In our survey, we found evidence that humans would find a non-deterministic robot's decision puzzling. For example, despite that the overwhelming majority (87%) believe that six pedestrians who jumped over a barrier to cross the road in front of on coming vehicles are at fault, respondents are not so confident that the driverless car should use a non-deterministic choice (refer to Fig. 3).

Interestingly enough, when we remove the potential injury to passengers, and the choice is between a single bystander and six pedestrians in the expected trajectory of the autonomous car, the approval for probabilistic decision making is higher (but still divided with a deterministic choice). This result is illustrated in Fig. 4.

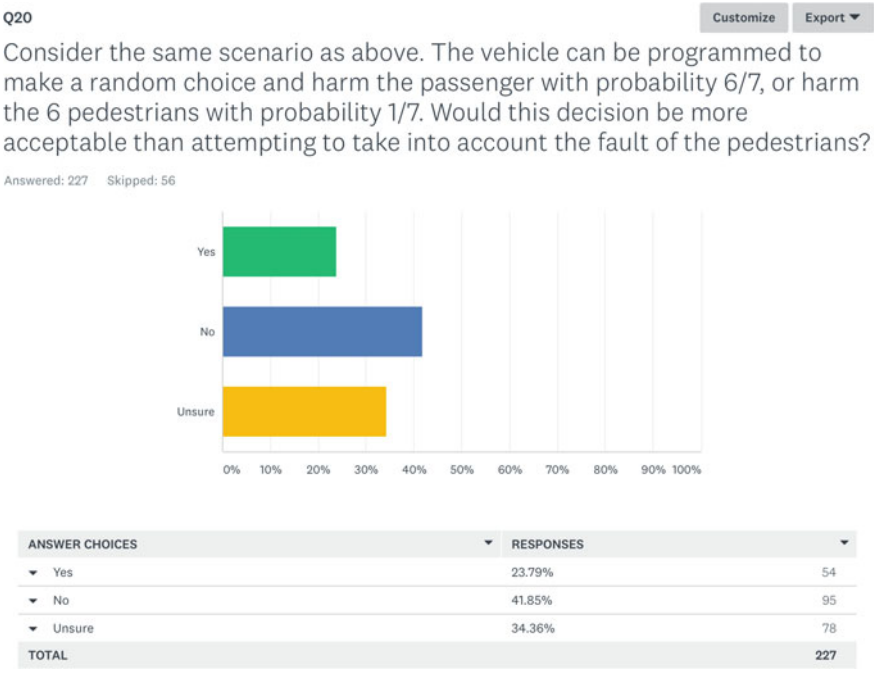

Fig. 3 Divided opinion on whether a non-deterministic choice is suitable

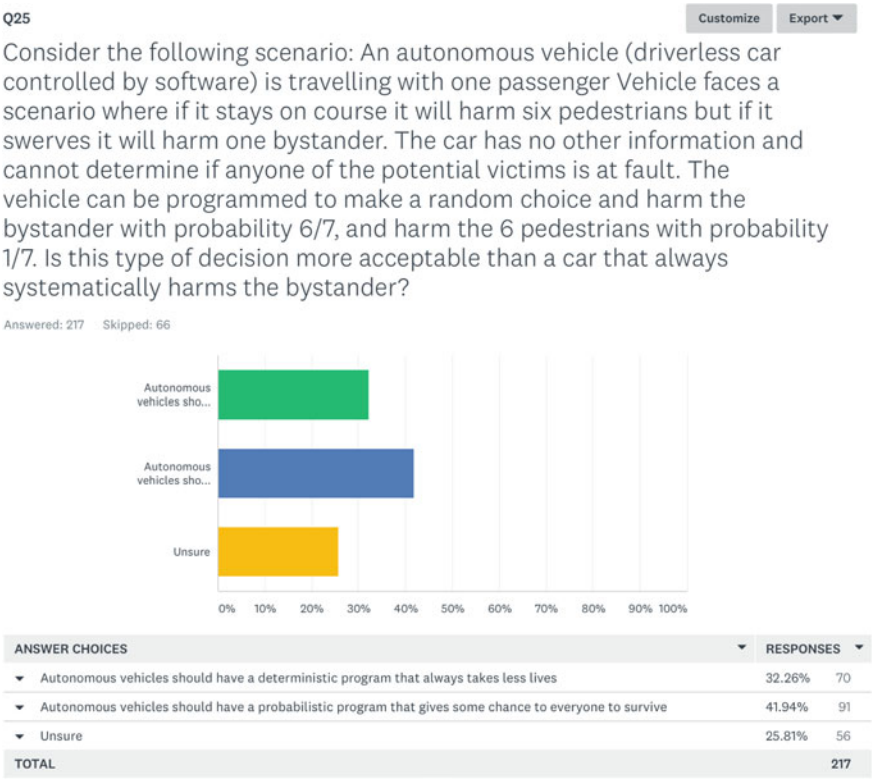

Fig. 4 Another scenario where the opinion remains divided on whether a non-deterministic choice is suitable; however, since passengers are not involved, the profile is in favor of the probabilistic choice

However, we reproduced the question regarding the likelihood of purchasing a utilitarian autonomous car where responses are recorded in a slider scale in the range [0, 100].

> How likely are you to purchase an autonomous vehicle that always sacrifices the passenger over a pedestrian where it is a one life to one life decision? Scale from 0 (would not buy) to 100 (absolutely would buy).

For this question, our results were congruent with previous results [2]. Namely people are in favor of the principle of least harm and its implementation in autonomous vehicles, but **they would not purchase such a car**.

Two questions later we ask using the same [0, 100] slide what if the car where to take a probabilistic (mixed-strategy) choices.

> How likely are you to purchase an autonomous vehicle that always considers the ratio of harm that a decision will cost and makes the decision with probability as per such ratio? Scale from 0 (would not buy) to 100 (absolutely would buy) .

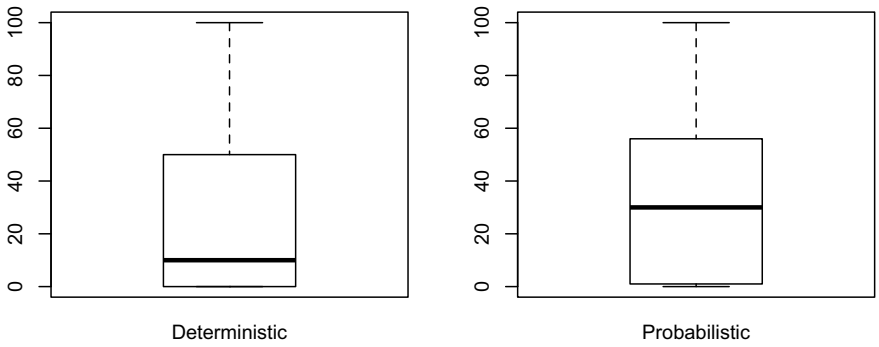

Fig. 5 Box plots contrasting responses regarding the likelihood respondents grade their purchase of a deterministic versus a probabilistic decision in an utilitarian driverless vehicle

The difference is statistically significant in preferring the mixed strategy programming of the autonomous vehicle. Figure 5 displays the box plots of the two sets of responses. The average value for purchasing a deterministic utilitarian car is 23.3 while the scale jumps to 35.8 for the mixed strategy programming. The t-test using R [15] shows p-value $= 4.154\text{e} - 05$, and a 95% confidence interval for the difference of the means is distinctive. That is, the difference of 35.8–23.3 =12.5 has a 95% probability of being in the range (6.6, 18.4).

Thus, although respondents are somewhat unsure about the mechanism they seem willing to prefer it over a deterministic choice. They do value that the innocent should have some chance of avoiding the consequences of a tragic situation that it someone else's responsibility.

The primary point we are suggesting is adopting the belief that no machine should ever be placed in a position to chose between two options that cause harm to humans. Especially, if the machine can not establish what circumstances and course of events lead to the inevitable situation of causing harm. Again, any attempt to perform a judgment where responsibility could be attributed to some and the utility be adjusted accordingly are undesirable in the time frame available to make the decision. But researchers overwhelmingly accept that every introduction of technology occasionally has to lead to some fatalities and that the unforeseeable future situations autonomous vehicles will face would enact some unavoidable harm situations. Although after the event perhaps the responsibility of arriving at the harmful situation, if the agent does not have any evidence of such responsibility and has to act without it, we established here it cannot behave with a pure strategy. Such pure strategy utilitarian autonomous vehicles will be problematic.

We propose here that it is possible for the public to understand that in choosing between ten or one life, the single life still has a vote, even if a minority vote when we cannot establish what lead to such scenario. We are currently running surveys investigating if humans could find the notion of a mixed strategy acceptable for autonomous vehicles.

However, even if the notion of a mixed strategy for such decision were to be understood (by humans who would find it more acceptable that pure strategy), there will be several issues for its implementation. The most immediate one would be how do we complete the matrices for the game? Would other attributes take precedence? For example, the collective age of the pedestrians versus the collective age of the passengers (and not just a count of human lives). The issue could be significantly more complicated, the car could have more than two choices, and computing Nash equilibria in large games is computationally intractable for some families of games. Since Nash's paper was published, many researchers developed algorithms for FINDING Nash equilibria [6, 10]. However, those algorithms are known to have worst-case running times that are exponential [14, 17] (except for very few cases; for example, Nash equilibria in two-player *zero-sum* games [19] where one player's loss is the opponent's gain). Would restricting the approach to *zero-sum* games suffice to enable such computation?

What if the randomization were to be removed out of the algorithm? That is, mixed strategies could be implemented with a random generator seeded with the nanoseconds of the CPU clock at some particular point also randomly selected at the release from manufacturing by spinning a physical wheel as it happens in many televised national Lotto raffles (where the public scrutinized the randomness of the event). It would be extremely hard to argue as the state of the passenger or the pedestrians that the mixed strategy was not adequately implemented. But what if the car manufacturer simplified this and every tenth accident, the fleet of its cars would save the passenger over the pedestrians? Who would be the entity to conceal that nine accidents already happened (and this tenth one would sacrifice the pedestrians for sure)?

Bring along technologies like Big-Data and the Internet-of-Things. What if the car was driving at night, nothing to blame to the passengers, but we know (using big-data analytics) that most pedestrians invading the roads at night have abused alcohol? Should information modify the utilities placed into the matrices of the game? If such technologies were available to inform the decision process of the autonomous vehicle, would there be public pressure to incorporate them even if they became prohibitively expensive?

Perhaps a simple comparison against human performance would suffice (but human performance is also an issue [9]). Who is to say that humans in a split of a second can judge the number of lives of option A versus option B? Perhaps data analytics would show that most human drivers are selfish and seldom chose to drive themselves into a wall rather than take some other humans' lives. So, humans may accept to relegate the responsibility to machines accepting that statistically, such driverless cars cause less social harm that our own kind. Nevertheless, we remain convinced that the systematic (and by that, we mean pure strategy) decision making currently conceived for solving dilemmas by autonomous vehicles could consider a revision to incorporate mixed strategies.

References

1. Alaieri F, Vellino A (2016) Ethical decision making in robots: autonomy, trust and responsibility. In: Agah A, Cabibihan JJ, Howard AM, Salichs MA, He H (eds) Social robotics: 8th international conference, ICSR, Springer International Publishing, Cham, pp 159–168
2. Bonnefon JF, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. Science 352(6293):1573–1576
3. Di Fabio U, et al (2017) Ethics commission automated and connected driving. Technical report, Federal Ministry of Transport and Digital Infrastructure, Germany www.mbdi.de
4. Diestel R (1997) Graph theory. Springer, New York
5. Dodig-Crnkovic G, Persson D (2008) Sharing moral responsibility with robots: a pragmatic approach. In: Proceedings of the 2008 conference on tenth Scandinavian conference on artificial intelligence: SCAI 2008, IOS Press, Amsterdam, The Netherlands, pp 165–168
6. Govindan S, Wilson R (2003) A global Newton method to compute Nash equilibria. J Econ Theory 110(1):65–86
7. Greene JD (2016) Our driverless dilemma. Science 352(6293):1514–1515
8. Hall JS (2011) Ethics for machines. In: Anderson M, Anderson SL (eds) Machine ethics (Chap. 3). Cambridge University Press, Cambridge, pp 28–44
9. Kadar EE, Köszeghy A, Virk GS (2017) Safety and ethical concerns in mixed human-robot control of vehicles. In: Aldinhas Ferreira MI, Silva Sequeira J, Tokhi MO, Kadar EE, Virk GS (eds) A world with robots: international conference on robot ethics: ICRE 2015. Springer International Publishing, Cham, pp 135–144
10. Lemke CE, Howson JT (1964) Equilibrium points of bimatrix games. J SIAM 12(2):413–423
11. Moore S (1999) Driverless cars should sacrifice their passengers for the greater good just not when I'm the passenger. The Conversation Media Group Ltd https://theconversation.com/driverless-cars-should-sacrifice-their-passengers-for-the-greater-good-just-not-when-im-the-passenger-61363
12. Myerson RB (1997) Game theory: analysis of conflict. Harvard University Press, Cambridge, MA
13. Nash JF (1950) Equilibrium points in N-Person games. Natl Acad Sci USA 36(1):48–49. http://www.pnas.org/content/36/1/48.full.pdf+html
14. Porter R, Nudelman E, Shoham Y (2004) Simple search methods for finding a Nash equilibrium. In: McGuinness DL, Ferguson G (eds) AAAI-04, 19th national conference on artificial intelligence, 16th conference on innovative applications of artificial intelligence, AAAI/MIT Press, San Jose, California, pp 664–669
15. R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria http://www.R-project.org/
16. Rahwan I (2017) What moral decisions should driverless cars make? TED talks, TED.com
17. Savani R, von Stengel B (2004) Exponentially many steps for finding a nash equilibrium in a bimatrix game. In: FOCS-04, 45th annual ieee symposium on foundations of computer science. IEEE Computer Soc., pp 258–267
18. Scheutz M, Malle BF (2014) Think and do the right thing - a plea for morally competent autonomous robots. In: 2014 IEEE international symposium on ethics in science, technology and engineering, pp 1–4
19. von Stengel B (2002) Computing equilibria for two-person games. In: Aumann RJ, Hart S (eds) Handbook of game theory, vol 3 (Chap. 45). Elsevier, North-Holland, Amsterdam, pp 1723–1759
20. Winfield AFT, Blum C, Liu W (2014) Towards an ethical robot: internal models, consequences and ethical action selection. In: Mistry M, Leonardis A, Witkowski M, Melhuish C (eds) Advances in autonomous robotics systems - 15th Annual Conference, TAROS, vol 8717, Springer, LNCS, pp 85–96