International Journal of Human-Computer Studies 130 (2019) 40–46

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Information, uncertainty and the manipulability of artificial intelligence autonomous vehicles systems

António Osório*,[a], Alberto Pinto[b]

[a] Universitat Rovira i Virgili, Dept. of Economics and CREIP, Spain
[b] Universidade do Porto, Dept. of Mathematics, Portugal

ARTICLE INFO

JEL classification:
D81
L62
O32

Keywords:
Artificial intelligence
Autonomous vehicles
Manipulation
Malicious behavior
Uncertainty

ABSTRACT

In an avoidable harmful situation, autonomous vehicles systems are expected to choose the course of action that causes the less damage to everybody. However, this behavioral protocol implies some predictability. In this context, we show that if the autonomous vehicle decision process is perfectly known then malicious, opportunistic, terrorist, criminal and non-civic individuals may have incentives to manipulate it. Consequently, some levels of uncertainty are necessary for the system to be manipulation proof. Uncertainty removes the misbehavior incentives because it increases the risk and likelihood of unsuccessful manipulation. However, uncertainty may also decrease the quality of the decision process with negative impact in terms of efficiency and welfare for the society. We also discuss other possible solutions to this problem.

1. Introduction

Autonomous vehicles navigate without human intervention through a system of mapping and sensorial technologies (e.g., sensors, radars, laser lights, GPS, odometry, computer vision, etc.) that can detect location and surroundings, and identify the appropriate paths between obstacles and signage (Mukhtar et al., 2015; Pendleton et al., 2017; Sun et al., 2006). The potential benefits of this technology are enormous. Autonomous vehicles are expected to reduce traffic collisions, improve traffic flow, mobility, relieve individuals from driving, decrease fuel consumption, facilitate transportation and businesses operations, among other (Clements and Kockelman, 2017; Gao et al., 2016; Mahmassani, 2016; Meyer and Deix, 2014; Speranza, 2018; Van Arem et al., 2006). In spite of the enormous potential benefits, there are a large number of unresolved safety, technology, ethical, social, political, regulatory and legal issues (Chen and Wang, 2005; Lindqvist and Neumann, 2017; Neumann, 2016; Parkinson et al., 2017; Petit and Shladover, 2015; Złotowski et al., 2017; among other).[1] However, as the autonomous vehicles system develops, most of these problems are expected to be solved because their nature is imminently technological.

In this paper, we discuss a related, but simpler problem, which is more difficult to deter and control. We refer to manipulation and opportunistic behavior that explores the predictability of the decision process of the autonomous vehicles system. Regarding this issue (Lin, 2016) points out (see also Schäffner, 2018):

> *"If the crash-avoidance system of a robot car is generally known, then other drivers may be tempted to "game" it, e.g., by cutting in front of it, knowing that the automated car will slow down or swerve to avoid an accident."*

On the contrary, to human manipulation by artificial intelligence systems (Bostrom and Yudkowsky, 2014; Brundage et al., 2018; Ricci et al., 2011; Russell et al., 2015), this type of manipulation and opportunistic behavior has not been sufficiently studied in the literature. However, it represents an enormous threat to the society and to the stability of the autonomous vehicles system, because it is easy to execute by malicious and opportunistic individuals, terrorists, criminals or people with no moral values. Moreover, since this type of situation does not rely so much on technology and other objective considerations, they are not easy to solve and deal with. In this paper, we try to make some progress in this direction.

Goodall (2014) presents one of the first detailed studies on how autonomous vehicles should behave in an unavoidable harmful situation. Shortly after, Bonnefon et al. (2015, 2016) performed a series of

* Corresponding author.
E-mail address: antonio.osoriodacosta@urv.cat (A. Osório).

[1] For instance, ethical hackers and researchers were able to hack the Tesla Model S, turning off the dashboard and various functions during driving. The same happened with the Jeep Cherokee.

https://doi.org/10.1016/j.ijhcs.2019.05.003
Received 18 June 2018; Received in revised form 14 February 2019; Accepted 13 May 2019
Available online 14 May 2019
1071-5819/

experiments suggesting that autonomous vehicles, as well as other artificial intelligence systems, are morally expected to choose the course of action that will cause less damage to everybody, i.e., the utilitarian course of action (Deng, 2015; Greene, 2016; Santoni de Sio, 2017). A series of subsequent contributions have discussed other related aspects like: people acceptance, incomplete information, ethical issues, and the complex and subjective cost/benefit analysis involving human lives (Etzioni and Etzioni, 2017; Goodall, 2016; Hevelke and Nida-Rümelin, 2015; Thornton et al., 2017; Trappl, 2016).

However, the predictability of the utilitarian approach opens the possibility to manipulation and opportunistic behavior. In this context, in addition to Lin (2016) mentioned above, Schäffner (2018) also raises concerns on how easy it is to abuse and manipulate the system, and question how can the autonomous vehicles system recognize the individuals' motives. In this paper, we study possible solutions to deal with this problem.

In order to better understand and illustrate in more detail the kind of manipulation and opportunistic behavior that is discussed in this paper, consider the following example. Suppose an autonomous vehicle with an elderly man inside. Suddenly and unexpectedly, a young man crosses in front of the vehicle, in such a way that the autonomous vehicle is unable to avoid an accident. In this context, the autonomous vehicle must take an action based on the potential damage to the individuals and associated probabilities (Goodall, 2014; Jacob et al., 1988; Pomerol, 1997; Santoni de Sio, 2017). All the rest being equal, suppose that hitting the young man is more damaging than deviating and getting off the road, and that the young man's life is worth more than the elderly man's life (Posner and Sunstein, 2005). In this context, the less damaging course of action, i.e., the utilitarian course of action, is to prioritize the young man life by getting off the road and causing potential damage to the old man. This decision seems the most correct and adequate in this situation. The problem is that this decision is predictable, and predictability opens the possibility to manipulation.

In other words, the accident in this example could have been caused intentionally by an individual with malicious intentions. In fact, given the predictability of the autonomous vehicle decision process, this individual risks nothing, because he/she knows that in those circumstances the system would give priority to his/her life.

This example extends easily to other situations involving malicious or opportunistic individuals, terrorists, criminals and people with no moral values. The details of the story may change, but the main idea remains the same.

In this paper, we generalize the previous example and consider ways to deal with this type of problems. We show that if the autonomous vehicle decision process is perfectly known, then malicious individuals may have incentives to manipulate it. Consequently, in order for the system to be immune to manipulation we need some degree of uncertainty, i.e., either noise in the decision process (internal uncertainty) or the knowledge about the specificities of the autonomous vehicle decision and evaluation processes are kept private (external uncertainty).

Uncertainty is important because it removes the misbehaving incentives by increasing the risk and the likelihood of unsuccessful manipulation. In this context, we show that sufficiently high levels of uncertainty are necessary for the system to be manipulation proof.

In this context, we also discuss the characteristics and disadvantages of the different types of uncertainty. For instance, internal uncertainty may reduce the quality of the decision process, which has implications in terms of efficiency, justice and social welfare, while external uncertainty can be learned by rational individuals from the history of past decisions, which again makes the system predictable and unable to solve the malicious pedestrian problem.

We also discuss our results along the Kerckhoffs (1883) principle, and consider the redesign of the autonomous vehicles system as an alternative solution to uncertainty. However, this solution might be difficult to implement. Finally, we note that the autonomous vehicles systems becomes manipulation proof if it prioritizes the lives and interests of their passengers, which might be a striking observation in the context of the utilitarian approach to the autonomous vehicles decision process.

The paper is organized as follows: Section 2 presents the malicious pedestrian model and notation, Sections 3 and 4 analyze the cases of no uncertainty and uncertainty, respectively, and Section 5 concludes.

2. The malicious pedestrian model and notation

In this section, we present the actions and expected utilities of the actors involved in the malicious pedestrian problem described in the introduction. The same framework can then be generalized to other situations of the same kind involving malicious, opportunistic and non-civic behavior, as well as terrorist or criminal activities, in the context of autonomous vehicles or artificial intelligence in general.

Let the subscript "i" denote the passengers traveling **inside** the autonomous vehicle and the subscript "o" denotes the pedestrians (or other individuals) **outside** the autonomous vehicle that are relevant to the problem. Pedestrians can have either "**good**" or "**bad** " intentions, denoted with the subscripts "o_g" and "o_b ", respectively, i.e., the subscript "o " can take the values "o_g " or "o_b".

A pedestrian with good intentions is somebody that respects the traffic and the rules of the autonomous vehicle system, and behaves in a social adequate manner. A pedestrian with bad intentions (or malicious pedestrian) is somebody that intentionally breaks the rules or take advantage of the autonomous vehicle system for own benefit, pleasure or simply to cause damage to others. The pedestrian with bad intentions is a metaphor for all possible situations involving malicious or opportunistic individuals, terrorists, criminals and people with no civic values that can fit in our framework.

Autonomous vehicles can be involved in several different types of crashes, accidents or failures. In our context, suppose that the autonomous vehicle is involved in an unavoidable accident situation and must decide between multiple potentially harmful courses of action. Note that in most cases—involving opportunistic individuals, terrorists, criminals or people with no civic values—the available courses of action are not necessarily harmful in physical terms, but in material, economical, legal or social terms. In order to simplify the analysis, suppose that there are only two possible courses of action:

Action I: gives priority to protect the lives and interests of the passengers traveling inside the autonomous vehicle. In this case, with probability p the autonomous vehicle is able to avoid collision and nobody is damaged or suffers any type of loss, and with probability $1-p$ the autonomous vehicle is not able to avoid collision and the pedestrians outside the autonomous vehicle suffer a damage or loss with value $d_o > 0$.

Action O: gives priority to protect the lives and interests of the pedestrians outside the autonomous vehicle. In this case, with probability q the autonomous vehicle is able to avoid collision and nobody is damaged or suffers any type of loss, and with probability $1-q$ the autonomous vehicle is not able to avoid collision and the passengers traveling inside the autonomous vehicle suffer a damage or loss with value $d_i > 0$.

When nobody is damaged, the involved parties (i.e., passengers and pedestrians) obtain the same constant utility $v \geq 0$ (later, for simplicity, we normalize this utility to $v = 0$). In this context, we can compute the expected utilities of the parties involved in our problem (Fargier and Sabbadin, 2005).

The **expected utility of the passengers** in Action I and Action O is:

$$E(u_i|I) = pv + (1-p)(-\alpha_i d_o), \tag{1}$$

and,

$$E(u_i|O) = qv + (1-q)(-d_i), \tag{2}$$

respectively, where $\alpha_i \in [0, 1]$ translates the damage/losses suffered by

the pedestrians outside the autonomous vehicle into disutility for the passengers traveling inside the autonomous vehicle. In other words, civic passengers traveling inside the autonomous vehicle derive no satisfaction from the damage/losses suffered by the pedestrians outside the autonomous vehicle. Therefore, $-\alpha_i d_o$ is the disutility of the passengers traveling inside the autonomous vehicle because of the damage and/or losses caused to the pedestrians. For instance, if $\alpha_i = 0$ the passengers derive no disutility from the damage or losses on the pedestrians, while if $\alpha_i = 1$, the damage or losses on the pedestrians are as if they happen to the passengers. In reality, people care—up to a certain extent—for others. For that reason it is natural to expect that α_i will take some intermediate value in the interval [0, 1].

The passengers expected utility is independent of whether the pedestrians have good or bad intentions, because it is difficult for passengers to identify ex-ante the pedestrians' true intentions. This identification difficulty is in the base of the malicious pedestrian problem. Otherwise, the autonomous vehicle system could discriminate between pedestrian types and choose different courses of action.

However, in our model, the pedestrians expected utility depends on whether they have good or bad intentions. The**expected utility of the pedestrians with good intentions** in Action *I* and Action *O* is:

$$E(u_{o_g}|I) = pv + (1-p)(-d_{o_g}), \tag{3}$$

and,

$$E(u_{o_g}|O) = qv + (1-q)(-\alpha_{o_g} d_i), \tag{4}$$

respectively, where $\alpha_{o_g} \in [0, 1]$ has the same interpretation and intuition as α_i presented above. It translates the damage/losses suffered by the passengers traveling inside the autonomous vehicle into disutility for the pedestrians outside the autonomous vehicle. In other words, $-\alpha_{o_g} d_i$ is the disutility suffered by civic (i.e., with good intentions) pedestrians due to the damage/losses caused on the passengers.

The **expected utility of the pedestrians with bad intentions** in Action *I* and Action *O* is:

$$E(u_{o_b}|I) = pv + (1-p)(-d_{o_b}), \tag{5}$$

and,

$$E(u_{o_b}|O) = qv + (1-q)u_{o_b}, \tag{6}$$

respectively. Note that in this case the expected utility of the pedestrians with bad intentions has a different structure, because these individuals derive some positive utility ($u_{o_b} > 0$) from causing damage/losses (e.g., physical, economical, destabilizing/playing with the system, etc.) to the passengers traveling inside the autonomous vehicle. This aspect distinguishes pedestrians with bad intentions from pedestrians with good intentions.

3. The case with no uncertainty - results

Let us consider that the goal of the autonomous vehicle system is to maximize the aggregated welfare of the individuals in the society, i.e., the passengers traveling inside the autonomous vehicle and the pedestrians outside the autonomous vehicle.

In this context, the system optimal decision must consider with equal weight the passengers and pedestrians interests, but ignores the interest of the pedestrians with bad intentions (and the implications of their actions on themselves), because of their subverted objectives. Otherwise, the consideration of the malicious pedestrians expected utility would induce noise and affect negatively the quality of the decision process, because it would become dependent on some arbitrary utility that these individuals derive from damaging others (i.e., the value u_{o_b}). In such context, in order to be even more precise, the decision process should also depend on the utility that passengers could derive from seeing pedestrians with bad intentions being punished by means of decisions contrary to their interests, and so on—a judgement that belongs to the courts and the legal system.

Consequently, in order to avoid the difficulty and subjectivity associated with these issues, the autonomous vehicle system must consider the interest of the passengers and pedestrians with good intentions only. Therefore, the total expected welfare of Action *I* is:

$$E(u|I) = E(u_i|I) + E(u_{o_g}|I) = 2pv - (1-p)(\alpha_i d_{o_g} + d_{o_g}),$$

which is obtained by adding expressions (1) and (3), where in the expression (1) we have set $d_o = d_{o_g}$. Similarly, the total expected welfare of Action *O* is:

$$E(u|O) = E(u_i|O) + E(u_{o_g}|O) = 2qv - (1-q)(d_i + \alpha_{o_g} d_i),$$

which is obtained by adding expressions (2) and (4). Therefore, Action *I* is optimal if the aggregate expected utility of Action *I* is higher than the aggregate expected utility of Action *O*, i.e., if $E(u|I) \geq E(u|O)$, and the opposite otherwise.

The following result formalizes this argument.

Lemma 1 (*decision process*). *The autonomous vehicle system optimal decision gives priority to the passengers interests relatively to the pedestrians interests (Action I) if:*

$$(1-q)d_i \geq (1-p)d_{o_g}, \tag{7}$$

and the opposite (Action O) otherwise.

In order to simplify and without great loss of generality, in inequality (7), we have normalized $v = 0$ and set $\alpha_{o_g} = \alpha_i = \alpha$, i.e., passengers and good pedestrians care equally (i.e., give the same weight) about the damage/losses on others—an assumption that is natural in our context.

Therefore, the optimal decision depends on the probabilities (i.e., p and q) and on the magnitude of the damage/losses in passengers and pedestrians (i.e., d_i and d_{o_g}, respectively). Then, depending on the circumstances, which are summarized by these parameters, the autonomous vehicle system will take the optimal decision (Roberts, 2008).

However, some circumstances make some decisions more likely than others. For that reason, rational pedestrians with bad intentions can manipulate or disturb the system if they are sure that the system will take the Action *O*, which occurs when in case of collusion the potential damage/losses to passengers are lower than to pedestrians, i.e., when the ratio d_i/d_{o_g} is low, and/or when the probability of damage/losses to pedestrians are lower than to passengers, i.e., when the ratio $(1-q)/(1-p)$ is low. In those circumstances, pedestrians with bad intentions are sure that the autonomous vehicle system will give priority to protecting their interests relatively to the interests of the passengers inside the autonomous vehicle. Therefore, pedestrians with bad intentions can manipulate the system without risking an adverse outcome, and they have the incentives to do it because their expected utility is positive, i.e., $E(u_{o_b}|O) > 0$.

The following result formalizes this argument.

Proposition 1 (*manipulation incentives with no uncertainty*). *If the malicious pedestrian is able to make correct judgments about the reality and has perfect information about the decision process of the autonomous vehicle system, then the autonomous vehicle system is manipulable.*

This result establishes the necessary conditions for the autonomous vehicle system to be manipulable. It does not mean that the system will be the object of manipulation, but rather that if pedestrians with bad intentions exist and the decision process is known, then the incentives to manipulate the system will also exist.

In order to complete our argument, we need to be more precise about what we mean by "perfect information" and "correct judgements about the reality". In our context, "perfect information" means that individuals know the algorithm underlying the decision process of the autonomous vehicle system. "Correct judgements about the reality" means that the individuals evaluation of the parameters p, q, d_i and d_{o_g} is in line with the evaluation made by the autonomous vehicle system (Jacob et al., 1988; Pomerol, 1997).

4. The case with uncertainty - results

Following the discussion, in order to remove the incentives to manipulate and to solve the malicious pedestrian problem, individuals with bad intentions must hold some uncertainty about the decision and evaluation processes of the autonomous vehicle system. Noise or observation difficulties reduce the incentives to misbehave. However, uncertainty may also reduce the quality of the decision process. In this section, we study these issues.

In order to capture the effects of uncertainty, we introduce uncertainty in the decision process, i.e., we add some noise component to the decision process characterized by inequality (7). In this context, let μ denote the probability with which the autonomous vehicle system gives priority to the passengers lives and/or interests (Action I), and $1-\mu$ denote the probability with which the autonomous vehicle system gives priority to the pedestrians lives and/or interests (Action O), where the probability μ is related with inequality (7) in a meaningful way with the addition of some noisy component (see Example 1 below for an illustration). This probability depends positively on p and d_i, but negatively on q and d_{o_g}. For instance, if $(1-q)d_i \geq (1-p)d_{o_g}$, i.e., inequality (7) is satisfied, then it is likely, but not for sure that the system will take Action I (i.e., in this case μ is high, but not equal to one). Similarly, if $(1-q)d_i < (1-p)d_{o_g}$, i.e., inequality (7) is not satisfied, then it is likely, but not for sure that the system will take Action O (i.e., in this case μ is low, but not equal to zero).

Note that μ is also the probability with which Action I is the correct decision and the probability with which Action O is the incorrect decision, while $1-\mu$ is the probability with which Action O is the correct decision and the probability with which Action I is the incorrect decision.

In this context, pedestrians with bad intentions will attempt to manipulate the system (i.e., to cause damage/losses on others or to destabilize the system) only if the likelihood of ending up damaging themselves is low. Therefore, pedestrians with bad intentions will have no incentives to misbehave if the expected utility from misbehaving is negative (or non-positive), i.e.,

$$\mu E(u_{o_b}|I) + (1-\mu)E(u_{o_b}|O) \leq 0, \tag{8}$$

where in the left-hand side, we have expressions (5) and (6), and in the right-hand side, we have the null normalized utility from not attempting to manipulate or destabilize the system (i.e., $v=0$). The following result formalizes our argument.

Proposition 2 (*no manipulation condition under uncertainty*). *The malicious pedestrian has no incentives to manipulate the autonomous vehicle system if:*

$$\mu \geq \frac{(1-q)u_{o_b}}{(1-q)u_{o_b} + (1-p)d_{o_b}}, \tag{9}$$

where μ denote the probability with which the autonomous vehicle system gives priority to the passengers' lives and/or interests.

The proof follows from the discussion and the inequality (8). Note also that in order to simplify and without great loss of generality, we have normalized $v=0$ and we have set $\alpha_{o_g} = \alpha_i = \alpha$.

Since the right-hand side of inequality (9) is a number smaller than one, pedestrians with bad intentions would not attempt to take advantage of the system if the probability with which the system gives priority to the passengers' lives and/or interests (i.e., μ) is sufficiently high.

In this context, note that the extreme solution in which the autonomous vehicle system gives priority to passengers' lives in all circumstances (i.e., chooses Action I always, which is equivalent to set $\mu=1$) trivially solves the problem. In this case, the inequality (8) becomes $E(u_{o_b}|I) = -(1-p)d_{o_b} < 0$, and consequently, pedestrians with bad intentions (i.e., malicious or opportunistic individuals, terrorists or criminals) would not attempt to manipulate or take advantage of the system.[2]

However, this solution has a problem because Action I is not always optimal, and it is easy to find situations in which Action O has more chances of resulting in no damage/losses to everybody than Action I. For instance, when the value of the pedestrians' lives and/or interests are high relative to the value of the passengers' lives and/or interests, i.e., when d_{o_g} is high relatively to d_i (or when q is high relative to p, see inequality (7)). The following result formalizes this discussion.

Corollary 1. *Action I, with probability $\mu=1$, solves the malicious pedestrian problem, but Action I is not always optimal.*

The proof of the first sentence follows from inequality (9). For the proof of the second sentence, it is enough to show that there are parameter values in which Action O is optimal, i.e., inequality (7) fails.

In general, since the probability μ depends on p, q, d_i and d_{o_g}, but also on some noisy component (i.e., the uncertainty component), it takes values lower than one. In this context, in order to understand the implications of uncertainty in the malicious pedestrian incentives and in the quality of the decision process, we distinguish two cases:

(i) In the case in which inequality (7) is not satisfied, **uncertainty is needed** in order to guarantee that malicious individuals have no incentives to manipulate the autonomous vehicle system. Uncertainty increases the value of μ, which in this case is the likelihood that the wrong Action I is taken. However, it is this potential mistake in the decision process (i.e., an unexpected decision contrary to the pedestrians' interests) that disciplines the malicious pedestrians. See Example 1 below, which illustrates the trade-offs between μ and uncertainty (i.e., σ).

(ii) In the case in which inequality (7) is satisfied, **uncertainty is not needed** because without uncertainty the system would optimally choose Action I with probability one, and a rational malicious pedestrian knows it. In this case, uncertainty is going to reduce the value of μ, which in this case is the likelihood that the correct Action I is taken.

Therefore, uncertainty reduces the incentives to misbehave, but also the quality of the decision process. However, in order to solve the malicious pedestrian problem under the utilitarian approach, we must accept some loss of quality in the decision process, even if very small.[3]

Proposition 3 (*uncertainty effects*). *(i) If inequality (7) is not satisfied, sufficient levels of uncertainty are required solve the malicious pedestrian problem. (ii) Otherwise, uncertainty only reduces the quality of the decision process.*

The proof follows from the previous discussion and the properties of μ discussed in the beginning of this section.

The following numerical example illustrates the arguments in Propositions 1–3.

Example 1. In order to introduce uncertainty in the malicious pedestrian problem in a meaningful way, we add a noise component

[2] The possibility that the system always takes Action I has been discussed in the literature (Bonnefon et al., 2016; Deng, 2015; Goodall, 2014; Greene, 2016). The idea is that since passengers own the autonomous vehicle, this should prioritize their lives and interests. However, the utilitarian approach seems to have higher support and is the one that makes more sense in the context of artificial intelligent systems.

[3] Note that in the case in which inequality (7) is not satisfied, it must be because d_{o_g} and/or q are large enough, and d_i and/or p are small enough, which implies that the right-hand side of inequality (9) will tend to be small (in particular, if we let $d_i = d_{o_g} = d_{o_b}$ and the value of u_{o_b} is not too large). Therefore, the value of μ required to satisfy inequality (9) when inequality (7) is not satisfied is much smaller than when inequality (7) is satisfied. See Example 1 below.

to the decision process in inequality (7), by defining the random variable $X = m + \sigma Z$, where $m = (1-q)d_i - (1-p)d_{o_g}$ is the mean, σ is the uncertainty parameter and Z follows the standard Gaussian distribution. Then, in line with inequality (7), the probability with which the autonomous vehicle system gives priority to protect the passengers' lives and interests (i.e., Action *I*) is given by:

$$\mu = P(X \geq 0) = \int_0^{\infty} e^{-\frac{(x-((1-q)d_i-(1-p)d_{o_g}))^2}{2\sigma^2}} / (\sqrt{2\pi}\sigma) dx,$$

and the opposite otherwise (i.e., Action *O*).

In order to simplify suppose that $d_i = d_{o_g} = d_{o_b} = u_{o_b} = 1$, and let us consider specific numerical values that illustrate the two cases stated in Proposition 3:

(i) If $p = 0.5$ and $q = 0.75$, inequality (7) is not satisfied (i.e., $0.25 < 0.5$), and Action *O* would have been chosen if there was no uncertainty. In this case, pedestrians with bad intentions could manipulate the autonomous vehicle system without risking their lives or interests. However, uncertainty about the autonomous vehicle decision (or evaluation) process changes this scenario in a positive way. For instance, if $\sigma = 1$ then $\mu = 0.401$ and inequality (9) would hold (i.e., $0.401 \geq 0.333$). However, note that uncertainty must be sufficiently high. For instance, if $\sigma = 0.5$ then $\mu = 0.309$ and inequality (9) would fail (i.e., $0.309 < 0.333$), and malicious pedestrians would have incentives to misbehave.

(ii) If $p = 0.75$ and $q = 0.5$, we have the case in which uncertainty does not help to solve the malicious pedestrian problem. In this case, inequality (7) is satisfied (i.e., $0.5 > 0.25$), and Action *I* would have been chosen if there was no uncertainty. Pedestrians with bad intentions would have no incentives to manipulate the autonomous vehicle system. The same happens if uncertainty is low. For instance, if $\sigma = 0.5$ then $\mu = 0.691$ and inequality (9) would hold ($0.691 \geq 0.667$), and malicious pedestrians would have no incentives to misbehave. However, high levels of uncertainty about the decision (or evaluation) process may change the things in a negative way. For instance, if $\sigma = 1$ then $\mu = 0.600$ and inequality (9) would fail (i.e., $0.600 < 0.667$).

The levels of uncertainty in the example are chosen to highlight the positive and negative effects of uncertainty.

The main observation of Proposition 3 and Example 1 is that uncertainty creates a trade-off—uncertainty is convenient in some circumstances, but not always. The problem is that in general, it is not possible to selectively have uncertainty. Uncertainty either exists or not, for good or bad.

When Action *O* is optimal in a deterministic context, uncertainty is convenient because allows the possibility that Action *I* is taken, which discipline the malicious pedestrian behavior. The disadvantage of uncertainty is that it may reduce the quality of the decision process, but not always. The reduction of quality depends on the structure of uncertainty. Consequently, we distinguish between two different structures of uncertainty:

Internal uncertainty corresponds to uncertainty in the decision process, which is then passed to the individuals. The system follows the decision rule in inequality (7), but with some noise, which creates uncertainty about how it functions. Consequently, individuals cannot anticipate perfectly what decision will be taken because the decision process is probabilistic. Uncertainty of this type has the disadvantage of reducing the quality of the decision process, and consequently the social welfare. In this context, we may observe situations in which the system mistakenly prioritizes the passengers' lives or interests and situations in which the system mistakenly prioritizes the pedestrians' lives or interests. In this paper, we do not explicitly model social welfare and justice, but mistaken decisions penalize unfavored individuals and introduce a loss of efficiency in the system and the society.

External uncertainty corresponds to uncertainty in the individuals about how the autonomous vehicle system takes decisions and evaluates the reality (e.g., how it determines the value of the parameters *p, q,* d_i and d_{o_g}). In other words, the system takes optimal deterministic decisions according to the decision rule in inequality (7), but are the individuals that do not know perfectly how these decisions are taken and how the system evaluates the reality (e.g., the algorithm behind these decisions). Since uncertainty is exclusively in the human side, there is no mistakes in the decision process. Uncertainty of this type does not reduce the quality of the decision process. However, in most contexts, it might be difficult to keep the decision process unknown for sufficiently large periods of time (Mercuri and Neumann, 2003), because rational individuals learn from observing the history of past decisions, and consequently are able to approximate its functioning, and find ways to successfully manipulate it.

The distinct characteristics of these two different structures of uncertainty may be important to the search and design of reliable solutions to the malicious pedestrian problem. However, some academics defend that security through uncertainty/obscurity is not appropriate at the system design stage. An old principle by Kerckhoffs (1883) argue that a well-designed security system, must not rely on secrecy. The "enemy" can learn the design without affecting the security.

In this context, an alternative solution to the malicious pedestrian problem would be to **redesign the autonomous vehicles system**, by restricting the way humans interact with the autonomous vehicles system. However, such possibility has implementation difficulties because it is costly and implies enormous changes in the design and architecture of the infrastructures, cities, roads and the society in general. In addition to these restrictions, the interaction between humans and the autonomous vehicles system is expected to be very diverse and free.

Consequently, we were forced to find security solutions in the decision algorithm behind the autonomous vehicles system, i.e., in the implementation stage. However, even at this stage, the uncertainty/obscurity approach is not consensual among academics (Almeshekah et al., 2013; Cowan, 2003; Swire, 2004; Witten et al., 2001). For instance, Hoepman and Jacobs (2008) argue that Kerckhoffs' Principle should also apply to the implementation stage, while Almeshekah et al. (2013) argue that uncertainty, obfuscation and deceptive information can be successfully applied to demotivates and slowdown malicious behavior.

In this context, Swire (2004) argue that the divergence of opinions between computer/networks experts and military/intelligence experts regarding the benefits of uncertainty/obscurity in security are due to the difference between the cyber and the physical world. Cyber-attacks are very cheap and malicious hackers can probe weaknesses in a software over and over again, while in the physical world, each attack is costly and the object under attack may adapt and change. This is an important aspect because the malicious pedestrian problem is a mixture of the cyber and the physical world, which involves human lives and interests, and exploits the predictability of the system.

We also note that these arguments and consideration are developed in the context of computer and network systems, and may not fit well into situations involving human lives, as the malicious pedestrian problem in this paper.

Given the lack of consensus and the limitations in the design of autonomous vehicles systems, a fourth alternative solution to the malicious pedestrian problem would be to **relax the utilitarian approach**. In this case, the autonomous vehicles system would protect the lives and interests of the passengers (i.e., Action *I* always), which is in line with the real world in that every being is self-interested and gives priority to their own lives and interests. This approach has the advantage of making the system manipulation proof (see Corollary 1), but it is not optimal and reduces the quality of the decision process. Moreover, it is contrary to the mainstream literature (see the discussion in the Introduction) that supports the utilitarian approach to the autonomous vehicles decision process (Bonnefon et al., 2016; Deng, 2015; Goodall, 2014; Greene, 2016; among others).

In terms of applied work, we note that in reality, the possible configurations of d_i, d_{o_g}, d_{o_b}, u_{o_b}, p and q are uncountable. In this context, we may have malicious pedestrians with very high values of u_{o_b}, for which malicious incentives can only be removed with very high levels of uncertainty. In those cases, given the negative implications that uncertainty has in the quality of the decision process and the low likelihood of those cases, the questions is whether it is worth to have uncertainty levels of such magnitude to dissuade these individuals from taking malicious or opportunistic actions. Considerations of this kind are important and should be taken into consideration in applied work.

Another important aspect in applied work is the complex and subjective task of instantaneous identification and calculations of the parameter values for each real life situation, i.e., the potential damage and associated probabilities. This is still an ongoing issue that requires further developments in mapping and sensorial technologies (e.g., sensors, radars, laser lights, GPS, odometry, computer vision, etc.). Consequently, the implemented algorithms and the chosen course of action will also depend on how precise and reliable these technologies will become (Mukhtar et al., 2015; Pendleton et al., 2017; Sun et al., 2006).

In practical terms, the objective of this paper is to call the attention of practitioners and the industry for the risks of manipulation and opportunistic behavior involving autonomous vehicles system and to discuss possible solutions to deal with this problem.

We conclude this section noting that malicious manipulation problems of the type discussed in this paper, which includes opportunistic, terrorist, criminal and other types of non-civic behavior, are not exclusive to autonomous vehicle systems, but likely to generalize to other artificial intelligence systems—with the necessary specificities and different levels of threat. The malicious pedestrian problem is a metaphor for a wider and general problem. In this context, our findings and arguments are generalizable to those systems, because as it is shown in this paper, the solution to this type of problems seems inevitable to require some degree of uncertainty in the perpetrator's side.

5. Conclusion

In this paper, we identify the existence of security concerns in autonomous vehicle systems caused by human actions (i.e., malicious, opportunistic, terrorist, criminal and non-civic actions) that explore the predictability of the decision process. We call to situations that have the same structure as the one described in this paper as "malicious pedestrian problems".

This type of manipulation has not received great attention in the literature,[4] but presents an enormous threat to the society and to the stability of the artificial intelligence systems. These problems are also more difficult to solve than other security problems (Chen and Wang, 2005; Lindqvist and Neumann, 2017; Neumann, 2016; Parkinson et al., 2017; Petit and Shladover, 2015; Złotowski et al., 2017; among other), because they do not rely on technological issues and other objective considerations. Simultaneously, they are very difficult to anticipate, detect or mitigate, and they are easy to execute because they require no skills and the physical access to the autonomous vehicles system network is public and available (Petit and Shladover, 2015).

We found that in order for the autonomous vehicle system to be immune to manipulation, individuals should not know perfectly how it functions. Otherwise, malicious manipulation is possible. In addition, we have discussed possible solutions to this problem. We found that uncertainty is a necessary condition for the system to be manipulation proof.

This observation implies that we should either (i) introduce uncertainty in the decision process (internal uncertainty) or (ii) keep the specificities of the decision process private (external uncertainty).

(i) The drawback of internal uncertainty is a reduction in the quality of the decision process and welfare. For that reason, it may not receive great support in contexts involving human lives, as in autonomous vehicle systems discussed in this paper. However, it might be the best solution in contexts in which individuals perfectly know or can easily learn the functioning of the decision process and in contexts in which there are no human lives involved (or other considerations of similar magnitude).

(ii) External uncertainty has the disadvantage that in most contexts, it is difficult to keep the decision process unknown for sufficiently large periods of time. This is the case because rational individuals can learn the functioning of the decision process from observing past decisions, and consequently approximate its functioning, and find ways to successfully manipulate it.

Another solution to the malicious pedestrian problem that avoids the resource to uncertainty, which is in line with the Kerckhoffs (1883) principle, is to restrict the way humans interact with the autonomous vehicles system. However, such possibility seems difficult to implement because it is costly and implies enormous changes in the design and architecture of the system infrastructures, cities, roads and the society in general.

Lastly, we note that the relaxation of the utilitarian approach can also be an alternative solution to the malicious pedestrian problem. In this case, the autonomous vehicles system would protect the lives and interests of the passengers, which is in line with the real world in that every being is self-interested and gives priority to their own lives and interests. This approach is manipulation proof, but it is not optimal and it is contrary to the utilitarian approach.

The results in this paper are shown in the context of the autonomous vehicle system, because autonomous vehicles are extremely intuitive, which makes the exposition simpler and easier to follow. Nonetheless, in one way or another, all artificial intelligence systems can be subject to human manipulation. This seems to be a limitation (or at least a weakness) of every artificial system. Even if the overall likelihood of occurrence and the implications of malicious, opportunistic, terrorist, criminal and non-civic behaviors is small, we cannot ignore the existence of these problems and their implications. This aspect is an important challenge to the future development of artificial intelligence systems (Kobbacy et al., 2007; Makridakis, 2017; Mingers and White, 2010). While, a large volume of literature has been studying how artificial intelligence can be used to influence and manipulate human behavior (Bostrom and Yudkowsky, 2014; Brundage et al., 2018; Ricci et al., 2011; Russell et al., 2015), the threats associated with human influence and manipulation on artificial intelligence systems opens an avenue for future research.

In this context, the crucial questions are how we will deal with these problems and whether the autonomous vehicles system will progress independently of their existence. We call for a research agenda on this type of problems.

Finally, we expect that our findings can help researchers and the industry choosing and designing the most effective manipulation proof artificial intelligence systems that can protect people and lead to better outcomes for the society as a whole.

Acknowledgments

We wish to thank to Juan Pablo Rincón-Zapatero, as well as several seminar and congress participants for helpful comments and discussions. Support from the Spanish Ministerio of Ciencia y Innovación project ECO2016-75410-P, GRODE Universitat Rovira i Virgili and Generalitat de Catalunya under projects 2018PFR-URV-B2-53 and 2017SGR770, LIAAD-INESC TEC, and the Portuguese Foundation for Science and Technology (FCT) projects PTDC/MAT-NAN/6890/2014

[4] The exception is Lin (2016) and Schäffner (2018), which point the possibility of manipulation and opportunistic behavior.

and PTDC/MAT-APL/31753/2017 are gratefully acknowledged.

References

Almeshekah, M.H., Spafford, E.H., Atallah, M.J., 2013. Improving security using deception. Center for Education and Research Information Assurance and Security, Purdue University, Tech. Rep. CERIAS Tech Report 13, 1–18.

Bonnefon, J.-F., Shariff, A., Rahwan, I., 2015. Autonomous vehicles need experimental ethics: are we ready for utilitarian cars? arXiv:1510.03346.

Bonnefon, J.-F., Shariff, A., Rahwan, I., 2016. The social dilemma of autonomous vehicles. Science 352 (6293), 1573–1576.

Bostrom, N., Yudkowsky, E., 2014. The ethics of artificial intelligence. Cambridge Handbook Artif.Intell. 316–334.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., et al., 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv:1802.07228.

Chen, H., Wang, F.-Y., 2005. Guest editors' introduction: artificial intelligence for homeland security. IEEE Intell. Syst. 20 (5), 12–16.

Clements, L.M., Kockelman, K.M., 2017. Economic effects of automated vehicles. Transp. Res. Rec. (2606), 106–114.

Cowan, C., 2003. Software security for open-source systems. IEEE Secur. Privacy 99 (1), 38–45.

Deng, B., 2015. The robot's dilemma. Nature 523 (7558), 24.

Etzioni, A., Etzioni, O., 2017. Incorporating ethics into artificial intelligence. J. Ethics 21 (4), 403–418.

Fargier, H., Sabbadin, R., 2005. Qualitative decision under uncertainty: back to expected utility. Artif. Intell. 164 (1–2), 245–280.

Gao, P., Kaas, H.-W., Mohr, D., Wee, D., 2016. Automotive revolution–perspective towards 2030 how the convergence of disruptive technology-driven trends could transform the auto industry. Adv. Ind. McKinsey Company.

Goodall, N., 2014. Ethical decision making during automated vehicle crashes. Transp. Res. Rec. (2424), 58–65.

Goodall, N.J., 2016. Can you program ethics into a self-driving car? IEEE Spectr. 53 (6), 28–58.

Greene, J.D., 2016. Our driverless dilemma. Science 352 (6293), 1514–1515.

Hevelke, A., Nida-Rümelin, J., 2015. Responsibility for crashes of autonomous vehicles: an ethical analysis. Sci.Eng.Ethics 21 (3), 619–630.

Hoepman, J.-H., Jacobs, B., 2008. Increased security through open source. arXiv:0801.3924. version: May 28, 2018

Jacob, V.S., Moore, J.C., Whinston, A.B., 1988. Artificial intelligence and the management science practitioner: rational choice and artificial intelligence. Interfaces 18 (4), 24–35.

Kerckhoffs, A., 1883. La cryptographie militaire, ou, Des chiffres usités en temps de guerre: avec un nouveau procédé de déchiffrement applicable aux systèmes à double clef. Librairie militaire de L. Baudoin.

Kobbacy, K.A., Vadera, S., Rasmy, M.H., 2007. AI And OR in management of operations: history and trends. J. Oper. Res. Soc. 58 (1), 10–28.

Lin, P., 2016. Why Ethics Matters for Autonomous Cars. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 69–85.

Lindqvist, U., Neumann, P.G., 2017. The future of the internet of things. Commun. ACM 60 (2), 26–30.

Mahmassani, H.S., 2016. 50th anniversary invited article–autonomous vehicles and connected vehicle systems: flow and operations considerations. Transp. Sci. 50 (4), 1140–1162.

Makridakis, S., 2017. The forthcoming artificial intelligence (AI) revolution: its impact on society and firms. Futures 90, 46–60.

Mercuri, R.T., Neumann, P.G., 2003. Security by obscurity. Commun. ACM 46 (11), 160.

Meyer, G., Deix, S., 2014. Research and Innovation for Automated Driving in Germany and Europe. Springer International Publishing, Cham, pp. 71–81.

Mingers, J., White, L., 2010. A review of the recent contribution of systems thinking to operational research and management science. Eur.J.Oper.Res. 207 (3), 1147–1161.

Mukhtar, A., Xia, L., Tang, T.B., 2015. Vehicle detection techniques for collision avoidance systems: a review. IEEE Trans. Intell. Transp. Syst. 16 (5), 2318–2338.

Neumann, P.G., 2016. Risks of automation: a cautionary total-system perspective of our cyberfuture. Commun. ACM 59 (10), 26–30.

Parkinson, S., Ward, P., Wilson, K., Miller, J., 2017. Cyber threats facing autonomous and connected vehicles: future challenges. IEEE Trans. Intell. Transp. Syst. 18 (11), 2898–2915.

Pendleton, S.D., Andersen, H., Du, X., Shen, X., Meghjani, M., Eng, Y.H., Rus, D., Ang, M.H., 2017. Perception, planning, control, and coordination for autonomous vehicles. Machines 5 (1), 6.

Petit, J., Shladover, S.E., 2015. Potential cyberattacks on automated vehicles. IEEE Trans. Intell. Transp.Syst. 16 (2), 546–556.

Pomerol, J.-C., 1997. Artificial intelligence and human decision making. Eur. J. Oper. Res. 99 (1), 3–25.

Posner, E.A., Sunstein, C.R., 2005. Dollars and death. Univ. Chicago Law Rev. 72 (2), 537–598.

Ricci, F., Rokach, L., Shapira, B., 2011. Introduction to recommender systems handbook. Recommender systems handbook. Springer, pp. 1–35.

Roberts, F.S., 2008. Computer science and decision theory. Annal. Oper. Res. 163 (1), 209.

Russell, S., Dewey, D., Tegmark, M., 2015. Research priorities for robust and beneficial artificial intelligence. AI Mag. 36 (4), 105–114.

Schäffner, V., 2018. Caught up in ethical dilemmas: an adapted consequentialist perspective on self-driving vehicles. Envisioning Robots in Society–Power, Politics, and Public Space: Proceedings of Robophilosophy 2018/TRANSOR 2018 Frontiers in Artificial Intelligence and Applications, 327–335.

Santoni de Sio, F., 2017. Killing by autonomous vehicles and the legal doctrine of necessity. Ethical Theory Moral Pract. 20 (2), 411–429.

Speranza, M.G., 2018. Trends in transportation and logistics. Eur. J. Oper. Res. 264 (3), 830–836.

Sun, Z., Bebis, G., Miller, R., 2006. On-road vehicle detection: a review. IEEE Trans.Pattern Anal.Mach.Intell. 28 (5), 694–711.

Swire, P.P., 2004. A model for when disclosure helps security: what is different about computer and network security. J. Telecommun.High Technol.Law 3 (1), 163–208.

Thornton, S.M., Pan, S., Erlien, S.M., Gerdes, J.C., 2017. Incorporating ethical considerations into automated vehicle control. IEEE Trans. Intell. Transp. Syst. 18 (6), 1429–1439.

Trappl, R., 2016. Ethical systems for self-driving cars: an introduction. Appl. Artif. Intell. 30 (8), 745–747.

Van Arem, B., Van Driel, C.J., Visser, R., 2006. The impact of cooperative adaptive cruise control on traffic-flow characteristics. IEEE Trans. Intell. Transp. Syst. 7 (4), 429–436.

Witten, B., Landwehr, C., Caloyannides, M., 2001. Does open source improve system security? IEEE Softw. 18 (5), 57–61.

Złotowski, J., Yogeeswaran, K., Bartneck, C., 2017. Can we control it? autonomous robots threaten human identity, uniqueness, safety, and resources. Int. J. Human-Comput. Stud. 100, 48–54.