Towards a formal ethics for autonomous cars

Michael P. Musielewicz Piotr Kulicki Robert Trypuz

The John Paul II Catholic University of Lublin,
Faculty of Philosophy, Al. Racławickie 14, 20-950 Lublin, Poland
{*michael.musielewicz,kulicki,trypuz*}*@kul.pl*

Abstract

Autonomous cars are one of the emerging technologies that will have a significant impact on society in the upcoming years. Although the predictions estimate that the traffic safety will be significantly improved, many people are afraid and prefer a human driver's control over vehicles or at least human driver's possibility to take control over the car. One of the reasons is that people want to be sure that in case of *hazardous situation or accident* a self-driving car will behave in a proper way. What does it mean "proper way"? There are several levels that can be considered, but at the end there is a level of values, especially moral values. In the paper we move towards a formal ethics for autonomous vehicles, which will allow people to understand the values influencing a self-driving car. To accomplish this we address philosophical concerns for the possibility of ethics for driverless cars, by paying particular attention to issue of their capacity as a normative agent. Moreover, we discuss a formal ontology for these vehicles and the possibility of the use of such ontology as a basis of a normative system. The lack of expressive power of ontological tools leads to the conclusion that the formal ethics for autonomous cars requires more powerful logic. The logic should be able to take into account norms on actions and states, and handle normative conflict and preferences on norms.

Keywords: driverless cars, formal ethics, normative agents, deontic reasoning, conflicts of norms, preferences

1 Introduction

Autonomous cars (also called driverless or self-driving) are one of the emerging technologies that is expected to have a significant impact on society in the upcoming years. The number of companies preparing to manufacture fully autonomous cars is growing and includes major car manufacturers, IT companies like Google, Intel or Nvidia, transport companies like Uber thereby making the economic and social expectations in that matter considerable.

Designers and producers of self-driving cars use a variety of technologies to develop the software for controlling vehicles. While there are particular differences, every one of them use some variant of statistical methods (applying such tools as artificial neural networks, deep learning, reinforcement learning etc.) as their main tool. These techniques are working very well in typical

situations, and at times can even outperform humans [1], however the process, leads to a black box algorithm of car control. The system acts successfully but we do not really know why and how (in the sense that the car's choices cannot be explained in a way that would be understood by an average person). That makes the technology suspicious for many parties involved in the use of autonomous cars.

Moreover, documents from regulatory authorities, *e.g.* [22,4,2], and researchers working in the field, *viz.* [6,10,5], recognize the need for ethical considerations concerning the behaviour of autonomous vehicles. However, they do not provide any complex and well defined theory in that matter. We also believe that an ethics for self-driving cars is indeed necessary. Moreover, we think that the ethics in consideration should be formalized. That allows for a precise expression of ethical intuitions, which is important for the success of a social debate in the subject and may also be useful for self-driving control software specification and development.

We begin the remaining part of the paper with a justification of the need of transparency of the decisions made by autonomous cars and a discussion of factors relevant for those decisions. Then, we approach the postulated formal ethics for autonomous cars. We address some philosophical concerns for the possibility of ethics for driverless cars. We pay particular attention to the issue of the capacity of a driverless car, or its controlling software, to be a normative agent. After this section the paper discusses an ontology for these vehicles which concludes with a proposal of a formal normative reasoning that will be useful for building a formal ethics for autonomous cars.

2 Social acceptance of driverless cars

2.1 Benefits and threats

There is widespread agreement that driverless cars, once introduced, are going to have a major impact upon society. Both the United States of America and the European Union have taken it upon themselves to be ready for this change. In the GEAR 2030 report [2] for the European Commission, a sketch of the impact that driverless cars are expected to have upon society is provided. Here, the expected impact ranges from a 90% reduction in human error related in road accidents to increased social mobility and even to a reduction of pollution in the environment [2, p.40]. Likewise, the US federal government sees safety as the paramount feature of this new technology and hopes to see a reduction of up to 94% of traffic accidents in the US, along with increased mobility for disabled persons [22, p.5].

Nonetheless, currently tested self-driving cars did not avoid serious collisions. Tesla's car in 2016 failed to detect a large white 18-wheel truck and trailer crossing the highway. The car drove full speed under the trailer, causing

[1] c.f. the outcomes of Microsoft's Bejing team in 2015 ImageNet Large Scale Visual Recognition Competition where they were able to have a 3.57% error rate surpassing human average error rate of 5% for the first time. [13, pp.223–5]

the collision that killed the 40-year-old behind the wheel in the Tesla. Recently, an autonomous Uber car killed a woman walking in the street in the Arizona [2]. We can see that the use of autonomous cars is not free from serious risks.

The European Commission's report mentions more "new challenges for regulators and policy makers concerning e.g. road safety, environmental, societal and ethical issues, cybersecurity protection of personal data, competitiveness and jobs, etc. which need to be addressed" [2, p.40]. Solving them is needed to build up the social acceptance of driverless cars.

A psychological factor also have to be considered. Although the predictions estimate that traffic safety will be significantly improved, many people are afraid and prefer human driver's control over vehicles or at least the possibility of a human driver to take the control over the car. Even specialists in the area remain skeptical about the technology they create. Raj Rajkumar, a leading expert on robotics, who cooperates with General Motors in the construction of autonomous cars, describes the current status of the technology in the following way:

> We as humans understand the situation. We are cognitive, sentient beings. We comprehend, we reason, and we take action. When you have automated vehicles, they are just programmed to do certain things for certain scenarios. [3]

So the users of autonomous vehicles want to know and understand (at some level of generality) how the vehicles are programmed to "do certain things for certain scenarios". They want to be sure that in case of a hazardous situation or an accident a self-driving car will behave in a proper way. Yet we must consider, what do we mean when we say "proper way"?

2.2 The need for transparency

In most cases it is possible to avoid damage to property, health, and the life of passengers and other participants of traffic. Moreover, it seems credible that a well trained algorithm will perform far better in driving than the average human driver or even a very good driver, and so it would seem that ethical considerations for driverless cars is relegated to only extreme situations. But this is not necessarily the case. The effects that these devices have upon their users may differ depending upon how its program is made or trained. The US Federal Government's policy for driverless cars indicates that, "even in instances in which no explicit ethical rule or preference is intended, the programming of an HAV [(highly automated vehicles)] may establish an implicit or inherent decision rule with significant ethical consequences"[22, p.26].

However, the very ascription of values to these objects, resting upon implicit ethical values, must be made clear so that all stakeholders can ensure

---

[2] See `https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe` (retrieved March 20, 2018)

[3] See `https://www.technologyreview.com/s/602492/what-to-know-before-you-get-in-a-self-driving-car/` (retrieved March 1, 2018).

that these "ethical judgments and decisions are made consciously and intentionally" [22, p.26]. This claim for transparency is mirrored in the report made by the ethics commission of the *Bundesministerium für Verkehr und digitale Infrastruktur* (hereinafter BMVI) made June of 2017. Here the BMVI underscores the importance of maintaining the autonomy of people in making ethical decisions and the prospect of some programmer or commission deciding how a driverless car should act on our behalf is problematic [4, p.16].

Hod Lipson and Melba Kurman write in their book *Driverless* [13], drivers make countless calculations and risk assessments of their behavior and of the road as it unfolds around them. When drivers are thrown into a situation where life is at risk they must react accordingly. Do they swerve right and hit a wall, or hit some other vehicle? When it is people making these choices there is an air of spontaneity which allows for us to forgive poor decisions, however the same does not apply for autonomous vehicles. As they say, "those of us fortunate enough never to have had a sever traffic accident have not had to perform the uncomfortable task of publicly articulating why we reacted the way we did when faced with an unavoidable traffic accident. Driverless cars stir up consternation since they force us to publicly reveal this calculation. Even more challenging, driverless cars will require that, as a society, we agree on a uniform set of ethical codes that will guide the decision-making process of artificial-intelligence software when faced with an emergency" [13, p.252]. But it is precisely this sort of "digging out" of our ethical calculations that will allow for transparency in this public debate.

We concur that it is crucial for autonomous vehicles' designers, and moreover for all stakeholders in these decisions, to make clear what hierarchy of values they embed in their vehicles. This clarification will enable the potential owners and users of self-driving cars, other traffic participants, the public in general, and regulatory authorities to accept or reject the underlying ethics in the vehicle's decision making algorithms before the wide scale usage of such vehicles.

In order to provide transparency for this issue, the thesis of this paper is two fold. The first is to establish driverless cars as being entities that are first capable of holding such ethical obligations. The second task is to provide a preliminary formal modelling of self-driving cars and their environment, and in particular a model that uses a formal normative reasoning, as it is a useful step towards making a clear specification of stakeholders expectations concerning the behaviour of autonomous vehicles leading to the aforementioned social consensus in that matter.

2.3 Possible factors influencing self-driving cars' expected conduct

What kind of factors should be taken into account when the 'ethical' behaviour of self-driving cars is considered? Let us refer to some statements that can illustrate the breadth of possibilities.

Patrick Lin argues that the chief safety feature of driverless cars, that is its "crash-optimization", implicitly means targeting which object to hit in order

to optimize a crash [12, pp. 72 – 73]. He notes that if we adopt a preference for minimizing harm to our property the car would need to target objects of a lesser weight than the vehicle; yet if we wish to minimize the harm to other people's property we ought to target an object of greater weight than the vehicle.

Michael Taylor from Car and Driver magazine, reported in [19] that according to Christoph von Hugo, Head of Active Safety in Mercedes-Benz Passenger Cars, all of Mercedes-Benz's future self-driving cars will *prioritize saving the people they carry*, which although they later retracted the statement, and indicated they would follow whatever the law proscribes, highlights the difficulties in pinning down the best response. [16]

In general, can or should an autonomous car value one life more than another on the basis of their relation to that car (value the passenger or owner over other persons), age, sex, status or by applying some other criteria? These difficulties in our (in)ability to choose who to save is seen in the often discussed trolley problems.

On this precise point there are many different points of view. Take for example the report made by the BMVI. There they lay forth 20 ethical rules for automated and connected vehicular traffic. In the 9^{th} rule they proscribe

> In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited. It is also prohibited to offset victims against one another. General programming to reduce the number of personal injuries may be justifiable. Those parties involved in the generation of mobility risks must not sacrifice non-involved parties. [4, p.11]

These are fairly strong claims and are further supported by first three articles of the *Grundgesetz für die Bundesrepublik Deutschland*, and raises questions if such "targeting" of objects that happen to be people could even be constitutionally permitted within Germany. These claims also seen in important associations in civil society. The IEEE (the Institute of Electrical and Electronics Engineers) also commit their members to these very same standards. Therefore, it would seem to answer our questions concerning whether an driverless car can value one life more than another.

Notwithstanding that apparent answer, there is more to the story than that. If we look at MIT's Moral Machine (`http://moralmachine.mit.edu`), we see that people do in fact have preferences and seem capable of choosing between two bad options; and when they are given a series of choices of how to act in various dilemmas general trends emerge. For an informal example we can see that enforcing the law, preferring women to men, humans to animals, fit people to fat people are some of the preferences that are noticeableA more formal example of this is also seen the work of Bonnefon et al. [9] where they noticed a strong preference for cars that minimize harm as such (*i.e.* by choosing self-sacrifice or the sacrifice of even loved ones) but it is conjoined with a general reluctance to buy such a car for themselves or even to have that sort of ethics enforced by legal means.

3 Ethics and selfdriving cars – foundational problems

3.1 What is meant by ethics?

Various regulatory institutions like the U.S. Federal Government in its policy [22, p.26], the European Union in its GEAR 2030 report [2, p.40] and press releases [3] in and the German ethics commission of the BMVI in their report [4], all emphatically assert the need for ethics for driverless cars. However, *there is no one clear understanding what is meant by ethics.* Rather, there are some common themes presented within these various works.

First and foremost, they note that ethics covers the decision making processes within the vehicle both in terms of legal and moral reasoning (or in some cases a conflation of these two modes of reasoning). Additionally there needs to be a balancing of moral actions, legal actions, and goal oriented actions made in light of moral and legal reasoning. Furthermore, it is clear that while ethics is important in all stages of automation, it is most important for when the vehicles are at higher stages of autonomy. In these stages, the human agent takes a lesser (to even non-existent) role in the operation of the device. The common convention that these various institutions use is the SAE (Society of Automotive Engineers) levels of automation. Notably the BMVI's document calls into question if removing the human factor is a good thing from an ethical perspective [4, p.20].

In addition to these regulatory texts various philosophers have also ventured some essential features of an ethics for autonomous vehicles. Neil McBride, for example, offers the *A.C.T.I.V.E.* (Autonomy, Community, Transparency, Identity, Value, and Empathy) formulation of ethics for autonomous cars, within which the rules that govern the both human – human and human – machine relationships need to be addressed [15]. Lin, believes that ethics has a crucial role in establishing what underlying values we ascribe to objects, which has a direct bearing on how the driverless car's crash optimization will function. Additionally, both McBride and Lin underscore the importance of broader ethical considerations for these new devices. One such example "conservative driving" where the autonomous car is overly cautious and other drivers will try to "'game' it, *e.g.*, by cutting it in front of it knowing that the automated car will slow down or swerve to avoid an accident" [12, p.51].

Another example, found in McBride, poses the question of what rules should govern the community – ranging from regulators, mechanics, the supply chain, etc. – that is formed to support these new device's. [15, p.182]

In all of these works, certain aspects emerge. We are living in an age where there new technologies are introducing new agents, and stakeholders, and while there is a host of benefits these changes it calls into question how we should act. Ethics, in this context, then seems to be the establishment of norms that govern the actions of and between these various agents, both in terms of the human – human relationship and the human – robot relationship (to borrow from McBride). It is this understanding of ethics that we wish to use for this paper. Nevertheless, there are still some serious philosophical questions that

first need to be addressed in order to build a proper ethics for these new devices.

3.2 Cars as normative agents

A poignant problem in designing an ethics for driverless cars is the establishment of these devices as normative agents that operate within a given normative system. If we are to do this there are several factors that need to be considered. First, we need to see that they are agents. Then if they are agents, we must see if they are normative agents. To establish that autonomous vehicles are normative agents requires first that they are agents that are capable of bearing norms as such and second they are placeable inside of a "normative system." This, however, is no small feat and will depend greatly upon one's conception for norms. It is only once we have established this, that the movement towards a formal ethics for driverless cars makes sense. For once we have the driverless car *qua* normative agent, we can flush out a subsequent ontology and normative reasoning kinds to model it. Although attributing normative agency to computer programs seems to be quite natural for computer science oriented logicians, for many legal theorists and philosophers (ethicist) it is still strange, so in this section we will argue for the aforementioned points.

To begin we need to establish that autonomous vehicles are in fact agents. There are various senses of agency that are used in various fields. In a plain sense, being and agent simply means having the capacity to act. There are, however, other more technical uses of the term. The most natural place to start is with a consideration of agency within computer science, where White and Chopra say (citing another author) that in this field an agent is "a piece of software that acts on behalf of its user and tries to meet certain objectives or complete tasks without any direct input or direct supervision from its user" [17, p.6].

Trypuz, in *Formal Ontology of Action*, furthers this definition and provides a good list of features that artificial agents have as found in the literature, having the following attributes: is autonomous, is situated (embodied in or inhabits an environment), is reactive – senses its environment and is responsive to changes in the environment, acts upon is environment, is proactive – has a set of goals or tasks, contains inner representations of itself and its world, is rational – "acts in its own best interest, given the beliefs that it has about the world", has the ability to perform domain-oriented reasoning, is a persistent (software) entity, and has social ability – interacts (negotiates and cooperates) with other agents (and possibly humans) via some king of agent-communication language: it engages in dialogues and negotiates and coordinates transfer of information [20, p.40]. Given these notions driverless cars seem to meet the well established criteria for being agents with the computer science community. Yet to be an agent is one thing to be a normative agent is quite another.

The problem of normative agency comes to the foreground when we reflect upon the nature of norms in themselves on a philosophical level. When we consider norms, as rules that govern the behavior of various agents, we notice that they have two related aspects. First they set the bounds of obligated, for-

bidden, or permitted actions and second these actions are ascribable to agents within the system, (*i.e.* a normative agent, who is beholden to these rules by virtue of being in the system).

To make these features clear let us consider an example provided by Ota Weinberger in *Law, Institution and Legal Politics: Fundamental Problems of Legal Theory and Social Philosophy*, where he offers an example of a game of chess to describe what he calls the institutional nature of "social normative systems".

> The rules of the game of chess are defined by its basic conditions: chessboard, figures, starting positions, rules of operation etc. We might ask whether these rules should be regarded as normative rules or as definitions. If they were mere definitions the person who does not adhere to the rules would not be seen as infringing the 'duty of the chess-player', but simply as not playing chess.[footnote omitted] It is true that nobody is obliged to play chess; the rules of chess apply to the players not as a system imposed by society but only as a result of a voluntary participation in the game; but they are relevant for the possibility of setting acts since they lay down a behaviour in accordance with a duty and define the class of possible results of the game: the game which is won (or lost) [24, p.193].

If two players sit down to play chess, they voluntarily enter into a sort of norm-governed activity constrained by the normative rules of the game. Their moves are permitted (such as a pawn may move two spaces in its first move), obliged (a pawn may only take other pieces that are in its diagonals and one space away), or forbidden (a pawn may not capture a unit directly in front of it) in respect to the rules of the game being played.

Weinberger expands this conception of normative systems as a game into broader considerations of law and into other norm based systems. For our purposes we can see how traffic fit within this framework. The driver (and drivers in general) are duty-bound to obey traffic norms, that is to say the drivers by the very act of driving become the "players", the traffic norms constitute the "rules of the game" that they are "playing", and the current state of affairs of the road are much like the game board. The key difference consisting in the complexity of the system, the content and number of norms (described in the previous section as combining moral, legal, and social rules) and the price of failures (in terms of tickets or even possible damage to persons and property). While this is quite clear, what remains a question is whether or not driverless cars are in fact agents, in a normative sense, that can fit into this system.

While the above schema is convenient in that it allows people who wish to study norms (or rights broadly construed) using more analytic tools, the topic of building an ethics for driverless cars poses a unique problem for it. Normally it is rather simple to use this framework when we apply it to various normative systems, whether in a game of chess or driving a car. The agents are well defined, and so are the rules, and problems are typically introduced when there are normative conflicts or moral dilemmas. When applying this

theory to driverless cars we first need to confront, the question "are they really normative agents?" How can they be bearers of rights in the broad sense? Or to use Hohfeldian terminology [8, p.30], can they be part of the duty – claim, privilege – non-claim, liability – power, and disability – immunity relationships?

To underscore this issue let us provide two examples. The first example is common place. A person is trying to cross the road at an uncontrolled intersection. The pedestrian has a claim to cross the road unimpeded which places a duty upon a driver to slow down and allow the pedestrian to cross. Is the driverless car beholden to the duty to "let pedestrians cross the street!"?

A more drastic example can also be taken from the well know trolley problem. An unmanned driverless car is going down the street and it is faced with a dilemma. Its breaks have failed, and now its controlling algorithm needs to make a choice of hitting a person in its lane or two people in the lane next to it. Now a dilemma is introduced if the car is beholden to the rule "Thou shalt not kill!" or say "Maximize the good!" or "Don't commit a forbidden act!" But is this the case? Against whom can the people in this perilous situation invoke a duty not to kill them? People in general? Sure, but in this case that's vapid. The programmer? Of course, but s/he programmed it not to hit people already. How about the driverless car itself? That's not clear. If the answer is no, then it would seem that the car does not have a duty to "not kill!" nor a duty to "maximize the good!" nor even a duty to " avoid committing a forbidden act!" that corresponds to any person's right in such a situation, in much the same way as we would not ascribe normative agency to a bull or a falling rock. Yet, if it lacks normative agency, then it falls outside of the normative system. Leaving us in a *de facto* situation were nothing is forbidden for it, and therefore everything is implicitly permitted. But surely that cannot be the case, can it?

If we are to avoid this we first need to dig even further into the theory of rights. There are presently two prominent theories of rights "will theory" and "interest theory" see e.g. [14, p.62]. These views of rights differ in what is required of an agent in order to ascribe to that agent rights as such, or in particular claims – duties, non-claims – privileges etc., and make them normative agents within a particular system of rights.

The key difference rests in the importance of the right bearers' interests and wills in the matter. For interest theorists, the bearers of these rights need only to be a beneficiary (or have some interest in the claim – duty etc. relationship), and the will is not needed. Will theorists, however, maintain that the bearer of these rights, need to be able to take an active role in the fulfillment of these rights, or put otherwise, be able to actualize them, to demand or to waive their right, and their interests need not be protected.

Both of these theories capture some of our basic intuitions on what rights are in relation to their bearers. Will theory maintains the idea that we are little sovereigns over our rights and can dispense or invoke them as we please. In interest theory we maintain the notion that rights ought to somehow be to our benefit. There is, however, a problem with will theory, namely the criterion

of the necessity of the will excludes certain classes from bearing rights, even among human beings, that intuitively should have rights. These classes of persons would include the unborn, infants, invalid, and the senile among others, who not having the capacity to use their will to demand or enforce their rights. For example they do not have the capacity to demand or waive claim to not be arbitrarily killed against some other person who is capable of fulfilling the adjoining duty in the Hohfeldian sense. So as they are incapable of having or exercising their wills they would then would have no rights, which is not the case.

This leads us to the consideration of interest theory. When considering driverless cars they certainly have interests when operating within the context of driving on the street. For example, they have an interest in crossing a busy intersection so the may have a claim of "right of way" against some other driver and that other driver has a duty to yield to the driverless car under that rule. Additionally, they may have in interest in being properly maintained and have a claim on their owner to service them. What is perhaps most important, is that given this conception of rights driverless cars may be bearers of rights (broadly conceived) and that right is granted by the normative system within which it is operating, and thereby are normative agents.

4 Towards an ontology-based normative reasoning for a self-driving car

In [1] we read that a logical-based control implemented in self-driving car would contribute to its *self-explanatory capacity* by which the author means a justification and explanation, in human understandable way, what the car "has done, is doing, and will be doing, and why". In the same paper it is emphasized that the self-driving car should have implemented an *operating-system-rooted ethical control* by which the author means "logics that are connected to the operating-system level of [...] cars, and that ensure these cars meet all of their moral and legal obligations, never do what is morally or legally forbidden, invariably steer clear of the invidious, and, when appropriate, perform what is supererogatory".

In section 4.1 we present, in our opinion, one of the most promising modeling framework that proposes a logic-based modeling for autonomous vehicles. It possesses the *self-explanatory capacity* but does not provide *operating-system-rooted ethical control*. In section 4.2 we discuss what should be added to the framework to fill that gap.

4.1 Advanced Driving Assistant System Ontologies

In this section we shall present and discuss an example of ontology-based normative reasoning for a self-driving car taken from the most resent research in the field. We shall refer to the works of Lihua Zhao and her colleagues [27,25,26] where they propose Advanced Driving Assistant System Ontologies (hereinafter ADAS Ontologies) and some interesting ideas how to combine ontologies with logical rules of reasoning expressed in the Semantic Web Rule Language (henceforward SWRL). The works, however, do not propose insights

on how to incorporate ethical rules into the system. But the frameworks they present is flexible enough to be a good starting point for our purpose.

The authors say that a self-driving car should be able to "infer driving behavior by processing the knowledge." [26, p.1427]. The knowledge they have in mind comes from a mapping of the sensor data (collected by the car in a real time) onto the categories of a machine-understandable ontology. So the car's "raw" perception data is transformed into ontological facts in the car's knowledge base. The knowledge base is a source of information that is used for making driving decisions. The car's knowledge in ADAS ranges from the spatio-temporal knowledge (maps, driving paths at given time), to knowledge about itself (its type, size, current speed, etc), knowledge about rules (traffic and others) to follow. That knowledge is to be considered when a decision is taken at a given time and position. For instance in real-time the self-driving car monitors its speed, so it can be said that it is aware of its speed. If it also knows the speed limit on the road it is currently driving on, then it can make a decision to accommodate the speed to the limit. It can also take into account the weather conditions or other conditions connected to traffic, so other rules (like safety rules) can have an impact on its behavior.

The SWRL rules they propose are formulated by means of categories taken from the ADAS Ontologies. The rules are conditional and trigger the execution of actions according to current car's conditions. It is worth noting that the driving decisions are on the level of basic driving actions such as *Stop*, *TurnRight*, or *Give Way*.

In [26] the authors propose 14 SWRL rules to model "Right-of-Way rules at uncontrolled intersections and on narrow two-way roads." They identified the following three situations that may occur: "Before an intersection: Give way or move forward in comply with Right-of-Way rules", "At an intersection: Stop and give way to the other cars when upcoming collisions are detected" (see formula (1)) and "On a two-way lane: Move to the left side and give way to the other cars coming from the opposite side of the two-way lane."

$$\begin{aligned} MyCar(?car1) \wedge isRunningOn(?car1, ?int) \wedge \\ Intersection(?int) \wedge collisionWarningWith(?car1, ?car2) \Rightarrow \\ Stop(?car1) \wedge giveWay(?car1, ?car2) \end{aligned} \tag{1}$$

These rules are triggered only when the self-driving car receives a collision warning signal from a collision detection system. It should be stressed that both detecting and preventing collisions belong to the most important category of the self-driving car's tasks. The SWRL rule reasoner performs reasoning on a fragment of the whole knowledge base containing the ontological description of the current driving situation.

Rules proposed in [26] – like (1) – although they do not contain deontic qualifications, they can express norms. (1) says that if my car is approaching an intersection and receives a collision warning with another car, then my car is obliged to stop and give a way to the other car. One should notice that

$Stop(?car1)$ in (1) means in fact that $?car1$ is classified as an object that should stop. The ontological modelling proposed in [26] contains no action tokens. This is because there is no need in that framework for modeling possible actions and describing possible choices in a given situation. It is assumed that after a collision warning there is a unique rule that, by taking into account the car's conditions, will property classify the car as being obliged to carry out an action of certain type.

4.2 Towards ethical control

Let us start with the following example and attempt of its modeling using the framework described above. A self-driving car is driving down a two-lane street, where there is an approaching truck in the opposing lane and a pedestrian on the sidewalk to the right. All of a sudden the truck swerves into the self-driving car's lane setting off the car's collision detection system. The driverless car must now prevent the collision by swerving right and out of the way onto the sidewalk, which triggers another collision warning. Now that it has two warnings, the car must decide whether is should collide with the truck and avoid hitting the pedestrian or swerve out of the way and avoid hitting the truck but then hit the pedestrian.

$$MyCar(?car1) \wedge collisionWarningWith(?car1, ?car2) \Rightarrow preventCollision(?car1, ?car2) \tag{2}$$

$$preventCollision(?car1, ?car2) \Rightarrow TurnRight(?car1) \tag{3}$$

$$TurnRight(?car1) \Rightarrow collisionWarningWith(?car1, ?person1) \tag{4}$$

$$MyCar(?car1) \wedge collisionWarningWith(?car1, ?person1) \Rightarrow preventCollision(?car1, ?person1) \tag{5}$$

$$preventCollision(?car1, ?person1) \Rightarrow TurnLeft(?car1) \tag{6}$$

After triggering all the rules one by one, i.e. after triggering the rule (6), the car will come back to the initial conditions and will start triggering the rules from the beginning, starting from (2). So in the consequence the car will keep turning left and right until it will eventually hit either the car $?car2$ or the person $?person1$. It is clear that "deliberating by doing" is not always the best option.

In section 2.3 we discussed possible factors that influence a self-driving cars' expected conduct. We pointed at different priorities that can have an impact on the car's actions. These priorities are both based on and justified by the values that the car's manufacturer or society has decided to implement in it. For instance we may prefer minimizing harm to our property more than

minimizing the harm to other people's property or we may prefer saving the people the car carries more than pedestrians etc. It is evident that the logical framework responsible for the self-driving car's ethical control should be able to express explicitly norms and priorities on them (see e.g. [7,23]).

A need for ethical reasoning often appears in situations where there are a few options that have to be judged in the light of ethical values and preferences. It means that we need more than rules of the form of "if condition A, then do B". A "higher level" deontic action logic reasoning (like those found in [11]) is needed that will constitute deliberation layer where possible actions could be identified, evaluated and finally the best option is chosen.

Then ethical reasoning requires taking into account both basic actions and their social interpretation. In the case of self-driving cars by basic actions we mean "turn left", "go straight","stop" etc. and their social interpretation in the particular conditions could be: "kill", "hurt", "protect", etc. Moreover, normative evaluation of actions themselves (like if they obey traffic regulations) and their results (like harm they make to the environment) should be considered [18,21].

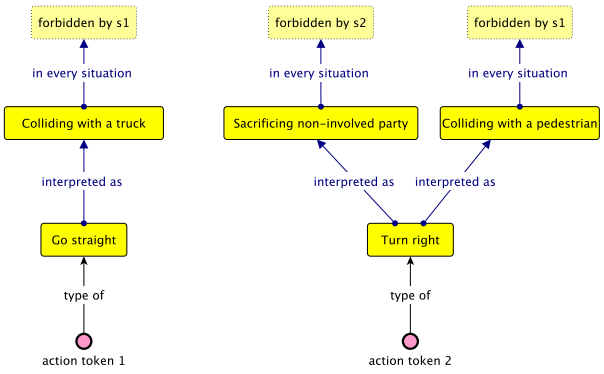

Figure 1. A situation where there are in force: a rule (from a normative source s1) that forbids collision with an object and a rule (from a normative source s2) stating that parties involved in the generation of mobility risks must not sacrifice non-involved parties. The self-driving car can carry out only two actions – action token 1 and action token 2.

For our example, if crash-optimization would be the main decision factor, then the choice between these options would be based on the evaluation of which object to hit in order to optimize the crash (the pedestrian has lesser weight than the car, contrary to the truck, so...). We could also discuss what would happen if the rule saying that "parties involved in the generation of mobility risks must not sacrifice non-involved parties" [4, p.11] would be implemented in the car. This rule could be of crucial importance and would forbid hitting the pedestrian that action would be recognized as "sacrificing a non-involved party". It is important here to interpret the car making the "turning right"

action as making the pedestrian involved. The former is not forbidden, while the latter is.

Following [11] we propose a modelling of the situation as depicted in figure 1. If we assume that the source of norms s2 is preferred over s1, then the deontic qualifications coming from s1 will be somehow "removed" from the normative system as being less important. Then we can conclude that the car should go straight. If neither of norms is preferred we face a dilemma situation and the procedure chosen for such a situation should be applied – in [11] several approaches for such a scenario are discussed. An implementation is proposed in a paper under review [4], however the programs in said paper are available online `http://kpi.kul.pl/deonticmachine`.

5 Conclusions

In this paper we have presented a justification for formal ethics for autonomous cars. The main point here is the need for transparency in car's behaviour that, as we believe, is a necessary precondition for the social acceptance of and the widespread introduction of this technology.

Here we have also discussed the some foundational problems of ethics for self-driving cars: the question what ethics means in this context and how cars can be understood as normative agents.

Finally we have examined the possibility of a "if-then rules" based approach to the specification of the expected behaviour of an autonomous vehicle. This approach, while useful, is not satisfactory when more complex situations are considered. That has led us the conclusion that more powerful logical tools are needed, and we have provided a list of the basic requirements of such a logic. Developing the particular details of this logical approach to the issue is planned as future work based on the findings of the present present paper.

References

[1] Bringsjord, S. and A. Sen, *On creative self-driving cars: Hire the computational logicians, fast*, Applied Artificial Intelligence **30** (2016), pp. 758–786, https://doi.org/10.1080/08839514.2016.1229906.

[2] European Commision, *The Report of the High Level Group on the Competitiveness and Sustainable Growth of the Automotive Industry in the European Union FINAL REPORT - 2017* (2017).

[3] European Parliament Press Room, *Robots: Legal Affairs Committee calls for EU-wide rules*, Press Release (2017), http://www.europarl.europa.eu/news/en/press-room/20170110IPR57613/robots-legal-affairs-committee-calls-for-eu-wide-rules.

[4] Federal Ministry of Transport and Digital Infrastructure, Ethics Commission, *Automated and Connected Driving* (2017).

[5] Gogoll, J. and J. F. Müller, *Autonomous cars: In favor of a mandatory ethics setting*, Science and Engineering Ethics **23** (2017), pp. 681–700.

[6] Goodall, N., *Ethical Decision Making During Automated Vehicle Crashes*, Transportation Research Record: Journal of the Transportation Research Board **2424** (2014).

[4] Submitted to Fundamenta Informaticae, 2017-07-07

[7] Hansen, J., *Deontic logics for prioritized imperatives*, Artif. Intell. Law **14** (2006), pp. 1–34, https://doi.org/10.1007/s10506-005-5081-x.
[8] Hohfeld, W., *Some fundamental legal conceptions as applied in judicial reasoning*, Yale Law Journal **23** (1913), http://digitalcommons.law.yale.edu/ylj/vol23/iss1/4.
[9] Jean-François Bonnefon, I. R., Azim Shariff, *The social silemma of autonomous vehicles*, Science **352** (2016), pp. 1573–1576.
[10] Johansson, R. and J. Nilsson, *Disarming the Trolley Problem –Why Self-driving Cars do not Need to Choose Whom to Kill*, in: M. Roy, editor, *Workshop CARS 2016 - Critical Automotive applications : Robustness & Safety*, CARS 2016 - Critical Automotive applications : Robustness & Safety, Göteborg, Sweden, 2016, https://hal.archives-ouvertes.fr/hal-01375606.
[11] Kulicki, P. and R. Trypuz, *Multivalued logics for conflicting norm*, in: *Deontic Logic and Normative Systems (DEON 2016)* (2016), pp. 123–138.
[12] Lin, P., "Autonomous Driving: Technical, Legal, and Social Aspects," Springer pp. 69 – 86.
[13] Lipson, H. and M. Kurman, "Driverless: Intelligent Cars and the Road Ahead (MIT Press)," The MIT Press, 2016.
[14] Matthew Kramer, H. S., N. E. Simmonds, "A Debate Over Rights: Philosophical Enquiries," 2000.
[15] McBride, N., *The ethics of driverless cars*, SIGCAS Computers & Society (2015).
[16] Orlove, R., *Now mercedes says its driverless cars won't run over pedestrians, that would be illegal*, Internet (2016), https://jalopnik.com/now-mercedes-says-its-driverless-cars-wont-run-over-ped-1787890432.
[17] Samar Chopra, L. F. W., "A legal Theory for Autonomous Artificial Agents," University of Michigan Press, 2011.
[18] Sergot, M., *Some examples formulated in a 'seeing to it that' logic: Illustrations, observations, problems*, in: *Outstanding Contributions to Logic*, Springer International Publishing, 2014 pp. 223–256.
[19] Taylor, M., *Self-driving mercedes-benzes will prioritize occupant safety over pedestrians*, Blog, https://blog.caranddriver.com/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/.
[20] Trypuz, R., "Formal Ontology of Action," Elpil, 2008.
[21] Trypuz, R. and P. Kulicki, *Connecting actions and states in deontic logic*, Studia Logica **105** (2017), pp. 915–942.
[22] U.S. Department of Transportation, National Highway Traffic Safety Administration, "Federal Automated Vehicle Policy Accelerating the Next Revolution in Road Safety," 2016, US Federal policy concerning AV.
[23] van Benthem, J., D. Grossi and F. Liu, *Priority structures in deontic logic*, Theoria **80** (2013), pp. 116–152.
[24] Weinberger, O., "Law, Institution and Legal Politics: Fundamental Problems of Legal Theory and Social Philosophy," Law and Philosophy Library 14, Springer Netherlands, 1991, 1 edition.
[25] Zhao, L., N. Arakawa, H. Wagatsuma and R. Ichise, *An ontology based map converter for intelligent vehicles*, in: T. Kawamura and H. Paulheim, editors, *Proceedings of the ISWC 2016 Posters & Demonstrations Track co-located with 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 19, 2016.*, CEUR Workshop Proceedings **1690** (2016), http://ceur-ws.org/Vol-1690/paper44.pdf.
[26] Zhao, L., R. Ichise, Z. Liu, S. Mita and Y. Sasaki, *Ontology-based driving decision making: A feasibility study at uncontrolled intersections*, IEICE Transactions **100-D** (2017), pp. 1425–1439, http://search.ieice.org/bin/summary.php?id=e100-d_7_1425.
[27] Zhao, L., R. Ichise, S. Mita and Y. Sasaki, *Core ontologies for safe autonomous driving*, in: S. Villata, J. Z. Pan and M. Dragoni, editors, *Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem, PA, USA, October 11, 2015.*, CEUR Workshop Proceedings **1486** (2015), http://ceur-ws.org/Vol-1486/paper_9.pdf.