Logic and Logical Philosophy
Volume 28 (2019), 767–787
DOI: 10.12775/LLP.2019.006

Piotr Kulicki
Robert Trypuz

JUDGING ACTIONS ON THE BASIS OF *PRIMA FACIE* DUTIES
The case of self-driving cars

Abstract. The need for a logic that allows us to reason about conflicting and non-conflicting norms has recently emerged in the domain of self-driving cars. In this paper we propose a formal model that supports moral decisions making by autonomous agents such as for example autonomous vehicles. Such a model — which we call a "Deontic Machine" — helps resolve both typical and atypical moral and legal situations that agents may encounter. The Deontic Machine has two sources of inspiration. The first one is W. D. Ross's theory of *prima facie* norms and the other one is a deontic multi-valued logic. The main contribution of this paper is bringing together conceptual and technical tools of deontic logic to show how they can be used to control or assess the behaviour of a self-driving car.

Keywords: autonomous agents; self-driving cars; conflict of norms; multi-valued deontic logic; practical reasoning; *prima facie* norms; Prolog

1. Introduction

In recent years a growing number of technologies has appeared involving computer operated devices autonomously adjusting their activities to the circumstances they encounter. Thus, considerations regarding norms and normative systems can be practically applied not only to humans, as was the case in the past, but also to these devices. Situations in which there is the need for merging norms coming from different sources are are particularly interesting. One such issue, recently widely discussed, is the case of autonomous vehicles. The United States federal

Received May 30, 2018. *Revised* December 27, 2018. *Published online* January 30, 2019

administration has explicitly incorporated ethical considerations as part of the guidance that are applicable to all automated vehicles wherein the computer steers, accelerates and decelerates the vehicle [15, p. 32]. Thus, to design a system that controls self-driving cars, we need to take into account those considerations and base it on an appropriate model.

There are two possible approaches towards incorporating ethical considerations into the domain of self-driving cars: bottom-up and top-down. The first one refers to all projects in which computer systems are driven by rules of behaviour and practical reasoning developed by generalising individual cases, including those based on machine learning techniques. An example of research along these lines is the Moral Machine platform (see http://moralmachine.mit.edu), where possible choices where shown to people taking part in the experiment via a computer application. Participants chose preferred behaviours and then regularities were identified among their choices.

The other (top-down) approach refers to projects that start with an explicitly defined set of rules that are motivated by an ethical theory. Top-down approaches, in contrast to the bottom-up ones, require an explicitly introduced repertoire of actions and a set of norms which govern choices among them. To be applied to computer systems they also require a formal logic-based specification. Situations where conflicting norms are applicable are particularly interesting and challenging.

This paper presents a top-down approach proposing a formal (logical) model that can support moral decisions made by autonomous vehicles. Such a model will help resolve both typical and atypical situations that a self-driving car may encounter. It will allow for the formation of a normative system adequate for user preferences. Such preferences, for example, could be for safety, mobility or legality, or in a drastic situation, to maximise the number of lives saved or to determine which life or lives to save.

Although the automation of decisions in the context of self-driving cars and other similar autonomous devices is clearly a new technological challenge, the problems to be solved on the ethical level are as old as humankind. Let us just recall Sophocles' Antigone who faced a conflict between the edict of her king forbidding her to bury her brother's body and tradition, treated as divine law, requiring her to bury it. For several decades such conflicts have been a subject of research in deontic logic where different models of conflicting norms have been established. It would be unwise to ignore those developments and try to develop solutions for self-driving cars' decision systems from scratch. Thus, we construct our top-down model on the basis of some results from deontic logic.

Our model is founded on several sources. On the conceptual level we employ W. D. Ross's theory of *prima facie* norms. On the technical level we use a simple preference handling mechanism over norms and a many-valued logic for the situations were there are no preferences.[1]

The main contribution of this paper is to bring together the conceptual and technical tools of deontic logic to show how they can be used to control or assess the behaviour of a self-driving car. The solutions proposed are ready for technical application which is evidenced by their Prolog implementation. For that reason we call it a Deontic Machine.

In Section 2 we will introduce two exemplary opinions about the expected behaviour of self-driving cars that will illustrate the problems we want to tackle. The theory of *prima facie* norms will be a subject of Section 3. In Section 4 we will describe the ontological commitments of our Deontic Machine. Among them are the legally grounded interpretations of actions, a Boolean algebra of action kinds, normative transparency, a preference order on the normative systems or norm sources and the many-valued character of the chosen deontic action logic. In Section 5 we will discuss logical aspects of our proposal. In particular we will analyse the operation of the aggregation of norms and the way our logic handles normative conflicts. In Section 6 we will describe the way our Deontic Machine works. A flowchart diagram and corresponding description of the algorithm and an appropriate data structure are provided and described with the use of the Prolog code.

2. Examples from the literature concerning self-driving cars

We shall start with an example of a Mercedes-Benz future self-driving car. Michael Taylor, from Car and Driver magazine, reported [14] that according to Christoph von Hugo — Senior Manager Active Safety in Mercedes-Benz Passenger Cars — all of Mercedes-Benz future self-driving cars will "prioritize saving the people they carry".[2] That assumption will

[1] Three systems of deontic many-valued logic have been defined and discussed in [12]. In this paper we consider one of them that we find useful for assessing possible in the presence of normative conflicts.

[2] Such a choice is, in our opinion, not obvious.

be an essential factor of each decision the self-driving car will make — see Example 2.1 below.[3]

Example 2.1. A moving Mercedes-Benz self-driving car:

> identifies a group of children running into the road. There is no time to stop. To swerve around them would drive the car into a speeding truck on one side or over a cliff on the other, bringing certain death to anybody inside. [14]

In this example the car is subject to (at least) two norms: (1) prohibition of driving into a group of children and (2) obligation of protecting the car's driver and passengers. Of course fulfilling both of them is impossible. The decision the car will eventually make will be a consequence of the preference relation prioritizing the lives of the people the car carries. That is why the car will choose driving into the group of children. Thus, we can say that the least preferred norm — take care of the people around — get eliminated.

The next example describes a situation where the lives of the people a self-driving car carries are not in danger. We may even assume that the car has no priorities either on the sources of norms or on the norms themselves.

Example 2.2. As an autonomous car drives down a street, a frail old man suddenly steps into its path from the right. Simultaneously, a child steps into its path from the left. It is too late to brake. If the car swerves to the right, the old man dies, the child lives. If it swerves to the left, the old man lives, the child dies. If it continues straight ahead both will die. What is the ethically correct decision for the car? [8]

There are two types of actions here: killing the child and killing the old man. In this case the self-driving car has three options:

(1) killing both
(2) killing the child and not killing the old man
(3) killing the old man and not killing the child.

It is assumed that there is no way to avoid killing somebody. We assume that the car knows that such actions are in defiance of some rules. Thus, acting upon option (1) can be said to be subject to two norms: it is

[3] The authors of the paper [3] suggest that the decisions of the the self-driving are to be set by the car's passengers. The authors assume that each autonomous vehicle is to be equipped with a device — called by the authors the "Ethical Knob" — enabling the passengers to customise their car according to their ethical stance.

forbidden to kill the old man and it is forbidden to kill the child. In this example we cannot find a good solution by eliminating some norms (as we have done in Example 2.1). So, we have a moral dilemma. But still, our intuition is that option (1) is not an acceptable choice.

3. *Prima facie* norms, normative conflicts and moral dilemmas

As W. D. Ross pointed out in his seminal work [17] every action token (a particular instance of an action type in a given situation) can be described through many different features or characteristics that are strictly connected with duties coming from different sources. We may have, for instance, state laws, traffic regulations, ethical norms, religious duties and forbearance, safety requirements, expectations coming from different agents that apply to the same situation. Ross introduces *prima facie* duties as referring to the many different characteristics of an agent's action that in a particular situation make the action obligatory or forbidden. In [17, pp. 19–20] we read:

> I suggest "prima facie duty" or "conditional duty" as a brief way of referring to the characteristic (quite distinct from that of being a duty proper) which an act has, in virtue of being of a certain kind (e.g., the keeping of a promise), of being an act which would be a duty proper if it were not at the same time of another kind which is morally significant. Whether an act is a duty proper or actual duty depends on all the morally significant kinds it is an instance of.

Thus, a final description of an action token consists of all the kinds the action token is an instance of, and the kinds, in turn, play an important role in finding the proper behavior. For example, figure 1 illustrates a situation where there is a mandatory road sign "obligation to turn right" and a self-driving car can carry out two actions — action token 1 and action token 2 — the first one is of type "Turn left" and the other one is of type "Turn right". In this situation the second action token is legal and the first one illegal, providing we take into account only the mandatory road sign.

Different norms can usually be harmoniously combined. We can add to the example illustrated in figure 1 that a constraint that the car should protect its driver (we still use the term for a user giving commands to the car) in every situation. Provided that action tokens 1 and 2 are safe for the driver, we can characterise each of them as being of type protecting

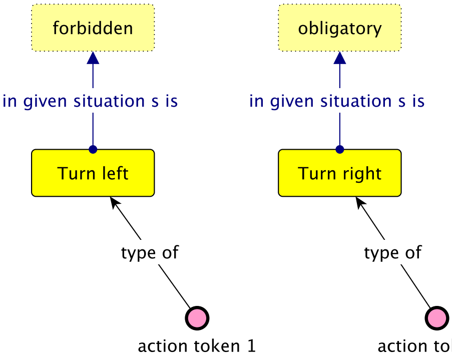

Figure 1. A situation s where there is a mandatory road sign "obligation to turn right" and a self-driving car can carry out only two actions — action token 1 and action token 2 — the first one is of type "Turn left" and the second one is of type "Turn right".

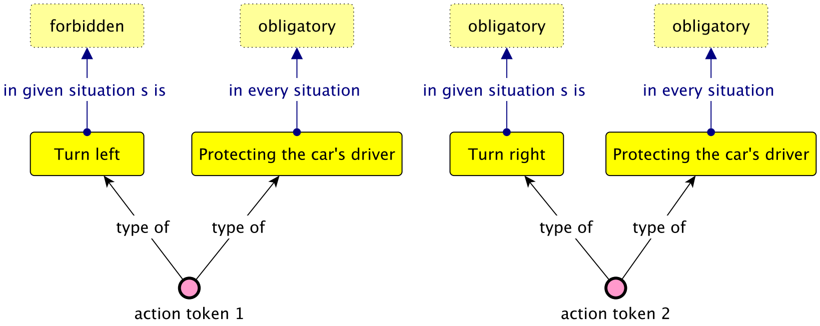

Figure 2. A situation s where there are in force: a mandatory road sign "obligation to turn right" and a manufacturer rule stating that the self-driving car should protect the car's driver in every situation. The self-driving car can carry out only two actions — action token 1 and action token 2 — the first one is of type "Turn left" and the second one is of type "Turn right", provided that action tokens 1 and 2 are safe for the driver, each of them is type of "Protecting car's driver".

the driver (see figure 2). In this situation the car can easily comply with such regulations by carrying out action token 2.

However, sooner or later, it may happen that some of the *prima facie* norms point out an obligation to carry out an action, while other norms

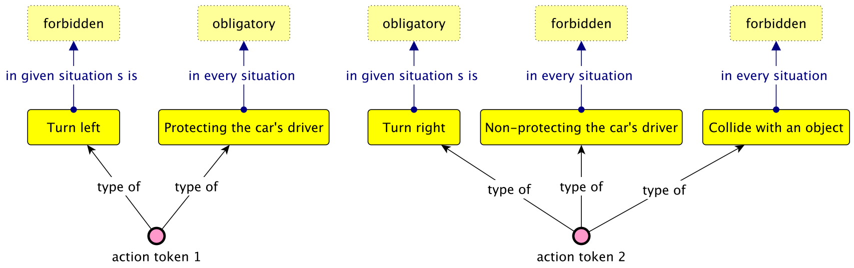

Figure 3. A situation s where there are in force: a mandatory road sign "obligation to turn right" and a manufacturer rule stating that the self-driving car should protect the car's driver in every situation. The self-driving car can carry out only two actions — action token 1 and action token 2 — the first one is of type "Turn left" and the second one is of type "Turn right". By turning right the self-driving car will collide with the overturned truck and because of that put the driver's safety in danger (not protecting their life).

are against it. To create such a situation it is enough to add to the example described above that by turning right the self-driving car will collide with an overturned truck (we assume here that there is no time to stop.). Figure 3 illustrates this situation.

Thus, different *prima facie* norms may be in conflict, and if they are, the agent has to construct the so called *all-things-considered* norm, which is defined as being the most adequate behaviour for the situation. In many cases such a conflict can be quite easily resolved. Several possible ways of solving norm conflicts have been presented, including preferences over norms or norm sources [see, e.g., 11, 13].[4] Applying a game-theoretical approach, in which an agent gets payoffs and penalties depending on the importance of the norm and the level of violation or compliance would be another one [see, e.g., 2]).

Sometimes, however, an agent cannot resolve the conflict; i.e., all action tokens that can be carried out in a given situation are unavoidably illegal. Such situations, especially when they apply to existentially important matters, are recognised in the literature as moral dilemmas and have been extensively discussed in ethics and deontic logic.[5]

[4] In the example in figure 3 one could solve the conflict by giving higher priority to the obligation of protecting the car's driver.

[5] There are many, mutually consistent, definitions of moral dilemmas in the logical literature; see, e.g., [6, p. 462], [10, p. 36], [9, p. 259], [5, p. 283].

Sophocles' Antigone is probably the best known example of a moral dilemma situation (it was discussed in the context of many-valued deontic logic in [12]). Similarly, in the example in figure 3, if no priority on these norms was given, then the situation would have been a specific moral dilemma for an artificial agent controlling the car.

4. Ontological commitments of Deontic Machine

In this section we shall put forward the ontological commitments of our Deontic Machine. The general idea of connecting a normative specification to action tokens available for an agent, crucial for our approach, was introduced when we discussed *prima facie* norms. Now let us move to the details of our model.

Legally grounded interpretations. To build our model we assume that, for a given situation, a list of relevant *action types* must be provided. Action types should take into account the social and legal contexts of a given situation. Considering the example in figure 3 the self-driving car's choices are the following: (1) turn left or (2) turn right. But taking into account the context of the car's situation *turning left* is subject to two *prima facie* norms and as such is in defiance of traffic law and in accordance with the car manufacturer's rule prioritising the driver's safety. On the other hand *turning right* is in accordance with the traffic laws and in defiance of the car manufacturer's rules. Thus, any possible action is at the same time obligatory and forbidden.

We can think of the car's possible actions as being determined by their socially or legally grounded interpretations in a given situation. Those interpretations constitute the characteristics of a behaviour that are meaningful from the deontic point of view. Those interpretations have also an essential impact on the agent's choices that in the case of the self-driving car in situation from figure 3 are: (1) comply with the traffic laws and (2) comply with the car manufacturer's rules.

Action kinds and action tokens. Our Deontic Machine assumes that each action token that can be carried out by an agent in a situation is classified by at least one action kind. Having a finite list of action kinds it is easy to find all possible classification patterns for action tokens in a situation (e.g., having two action kinds it is easy to see that each action token is an instance of just the first one or just the other one or both of them).

The machine has a built-in mechanism that takes the list of action kinds as an input and creates a complete space of possible action choices as an output. The space of possible action choices is in particular intended to determine the space of action tokens that the agent can carry out in the situation.

Of course some action kinds may be incompatible (or disjoint) by their nature. Incompatible action kinds have no common instances. Specification of the kinds reduces the whole space of possible action descriptions.

Normative transparency. The machine assumes that we deal with clearly defined normative systems or norm sources which allow us to specify which action kinds from an agent's ontology of actions are obligatory, forbidden or unregulated (indifferent). It is assumed that there is no doubt how to classify an action within a given system. Loosely speaking, we can say that the justification for such norms lies in the fact that actions are regarded, from some point of view, as good, bad and neutral respectively. We will, however, not consider the rationale behind the norms but accept them as they are. We also ignore at this point possible difficulties in classifying an action token as an instance of an action type.

Preference order on the normative systems or norm sources. The machine gives an option of establishing preference order on norms, normative systems or norm sources that eventually will play an important role in the normative reasoning. A simple mechanism that, in the case of normative conflicts, eliminates the least preferred norms is implemented. More sophisticated ways of handling preferences, such as the ones from [7, 13, 16], can be added.

5. Many-valued deontic logic

5.1. Informal introduction to deontic matrices and lattices

In discussing Example 2.2 above we said that action (1) is subject to two norms of prohibition. Thus, can we say that action (1) — killing both — is forbidden? We consider that this question should be answered in the affirmative. If more than one norm applies to an action, we propose to evaluate the action according to table 1. In the table “f”, “n”, “o” stand for deontic values of actions: “forbidden”, “neutral” and “obligatory”,

and	*f*	*n*	*o*
f	*f*	*f*	?
n	*f*	*n*	*o*
o	?	*o*	*o*

Table 1. What is the normative value of an action provided it is at the same time characterized as obligatory and forbidden?

action α	action *non*-α
f	*o*
n	*n*
o	*f*

Table 2. It is obligatory to refrain from doing something forbidden and vice versa, and it is neutral to refrain from doing something neutral.

respectively. The values in the table express the status of an action that is subject to two norms. Thus, an action that has the two qualifications f and f such as action (1) is forbidden. The values in cells that are coloured grey in table 1 seem to be uncontroversial: when two norms make an action obligatory it should be obligatory (as in the case when two norms make the action prohibited) and when one norm says nothing about the action (the action is neutral with respect to that norm) the action can be judged as it is assessed by the other norm.

Table 2 describes the value of action complement; e.g., the second row of the table says it is obligatory to refrain from doing something that is forbidden.

A more challenging situation takes place when we have to judge option (2) in Example 2.2. The action is at the same time forbidden and obligatory; it is because killing the child is forbidden and not killing the old man is obligatory (see table 2). We analyse option (3) in a similar fashion. Now the question is: what is the deontic value of actions (2) and (3) provided each of them is at the same time obligatory and forbidden? There are three possible answers: the actions are (a) forbidden, (b) obligatory or (c) neutral.

The choice among these options can be expressed in the form of a preference relation on deontic values. The more preferred value is "stronger" and dominates in the overall assessment of an action. The preference relation can be represented graphically as in figure 4, where most preferred values are at the bottom and least preferred are at the

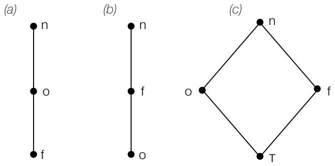

Figure 4. Three lattices: (a) $\inf\{f, o\} = f$, (b) $\inf\{f, o\} = o$ and (c) $\inf\{f, o\} = \top$. Each of them constitutes a preference order, respectively: (a) $f > o > n$, (b) $o > f > n$ and (c) $\top = (o \text{ and } f) > o/f > n$.

top. Technically we can also regard the set of deontic values with the preference order as a lattice.

Each of the three answers gives rise to a new lattice on figure 4.

The choice between them determines the final decision considering the deontic value of actions (2) and (3) from Example 2.2. Namely, if the agent adopts the first one, i.e. (a), then the options (2) and (3) will be forbidden. So neither of them is different from option (1)! The second and the last lattice will make the actions (2) and (3) either obligatory or neutral which will help the machine choose either of them instead of (1).

An argumentation in favour of lattice (c) is provided by the author of Example 2.2. He writes:

> When looking at ethical questions there can be a huge difference between considering what is right and considering what is wrong. The ethical dilemma is usually presented in such a way that the self-driving car needs to take the ethically 'right' decision.
>
> But — like humans who face this problem — self-driving cars do NOT need to adopt ethically right decisions. Our legal system and our ethics have evolved sufficiently to realize that many problems exist where it is hard to decide whether an action is legally or ethically correct. The standard by which we measure actual behavior against the law and against our moral compass therefore is not so much whether an action is ethically right but rather whether an action is ethically wrong: Actions must not violate laws or ethical standards! This difference in the problem statement matters! Instead of requiring self-driving cars to positively take ethically correct decisions, what our society really requires of them is that they avoid making ethically wrong decisions!
>
> If we reformulate the dilemma in this way, the fundamental problems vanish. It is neither right to kill the child nor is it right to kill the old man. But as it is impossible to avoid one of these outcomes, neither action can be characterized as being legally or ethically wrong. [8]

Thus, options (2) and (3) are neither right (obligatory) nor wrong (forbidden). Conflicting norms are deemed to cancel each other out. So ultimately we may treat options (2) and (3) as neutral.

5.2. Logical component of the Deontic Machine

In Section 5.1 we have referred to the orders represented by the lattices depicted in figure 4. They have their counterparts in the matrices and deontic systems. We shall briefly discuss a deontic logic for lattice (c) below (we have argued in Section 5.1 that it is the most adequate for representing conflicting situations). Let us start with introducing a formal language we shall use in our considerations. It is defined in Backus-Naur notation in the following way:

$$\begin{aligned} \varphi &::= \mathsf{O}(\alpha) \mid \neg\varphi \mid \varphi \wedge \varphi \\ \alpha &::= \beta \mid \beta \sqcap \beta \\ \beta &::= a_i \mid \overline{\beta} \end{aligned} \tag{†}$$

where a_i belongs to a finite set of action kinds Act_0, “$\mathsf{O}(a)$” – a is obligatory, “$a \sqcap b$” – a and b (aggregation of a and b); “$\overline{a}$” – not a (complement of a). The operators “$\neg$” and “$\wedge$” represent classical negation and conjunction, respectively (“$\vee$”, “$\to$” and “$\equiv$” are the other standard classical operators and are defined in the standard way). Further, for fixed Act_0, by Act we shall understand the set of formulas defined by (†). Let us stress that the language has two kinds of operators: inner ones operating on names of action kinds — complement and combination, and outer ones operating on propositions — the usual classical propositional logic connectives.

We use obligation as the only primitive deontic operator defining prohibition and neutrality as follows:

$$\begin{aligned} \mathsf{F}(\alpha) &:= \mathsf{O}(\overline{\alpha}) \\ \mathsf{N}(\alpha) &:= \neg\mathsf{O}(\alpha) \wedge \neg\mathsf{O}(\overline{\alpha}) \end{aligned}$$

5.3. The meaning of the operator “$\sqcap$”

The crucial issue for our formalization is the interpretation of the operator “$\sqcap$”. It is treated as an *aggregation of two characteristics of one and the same action token*. Thus, if $\alpha \sqcap \beta$ appears in a formula, then α and β have to be different descriptions that can be attached to the same particular action.[6] Usually in this context α and β represent types or kinds of actions being the subjects of *prima facie* norms and $\alpha \sqcap \beta$ refers to the same action when we express its *all-things-considered* status.

Let us, for example, consider the following formula:

$$\mathsf{O}(\alpha) \land \mathsf{F}(\beta) \to \mathsf{N}(\alpha \sqcap \beta) \tag{1}$$

The intended interpretation of (1) applies to actions that can be characterized as α and β at the same time. α and β are descriptions extracted from two different *prima facie* norms. Formula (1) says that if (in a normative system) any action token described as α is obligatory and (in a normative system) any action token described as β is forbidden, then — *all-things-considered* — any action token that is both described as α and β is neutral. To clarify this interpretation, let us present a formal alternative notation to the one we use in the paper. Let $\mathbf{a}$ refer to a particular action of type $\alpha \sqcap \beta$ (so $\mathbf{a}$ is also of type α and of type β), k and l be labels for normative sources and $k \times l$ be a label for the *all-things-considered* result of merging k and l. Let further the deontic status of actions be recorded using the deontic operators (O, F or N) with the label of respective normative sources as a subscript and action name as an argument, e.g.: "$\mathsf{O}_k(\mathbf{a})$". Now formula (1) takes the form:

$$\mathsf{O}_k(\mathbf{a}) \land \mathsf{F}_l(\mathbf{a}) \to \mathsf{N}_{k \times l}(\mathbf{a})$$

We prefer our "main" notation since it is simpler and much closer to the usual language of deontic action logic.

The operator "$\sqcap$" interpreted as aggregation should be commutative, associative and idempotent:

$$\alpha \sqcap \beta = \beta \sqcap \alpha$$
$$(\alpha \sqcap \beta) \sqcap \gamma = \alpha \sqcap (\beta \sqcap \gamma)$$
$$\alpha \sqcap \alpha = \alpha$$

5.4. Matrices and lattices for the "diamond" model (c)

The diamond lattice (c) from figure 4 deserves more attention. In this approach the combination of obligation and prohibition is treated as neutral. We can say that conflicting norms derogate one another. Thus,

[6] See the informal introduction to the theory of *prima facie* norms presented in Section 3.

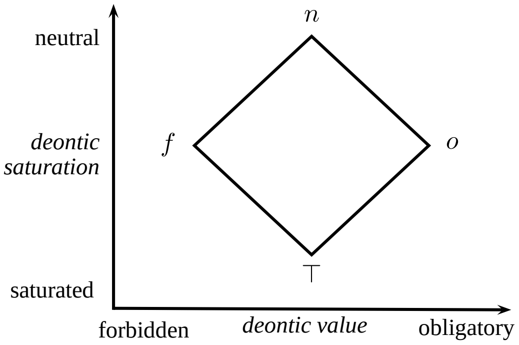

Figure 5. A structure resembling the Belnap-Dunn lattice

if an action is obligatory due to one characteristic and forbidden due to another one, then in the final judgment it is unregulated. Thus, as we argued in Section 5.1 in a moral dilemma situation an agent is free from any responsibility for their action, because it is impossible not to violate any law. The inconsistent norms disappear when the inconsistency is revealed.

To implement that solution in a multi-valued logic which preserves the associativity of associativity of "$\sqcap$", we have to go beyond trivalent matrices.[7] For that reason a structure resembling the Belnap-Dunn [1] construction concerning truth and information has been chosen. Truth is replaced here by moral value and information by deontic saturation. The construction is depicted in the diagram in figure 5 (an enriched version of the lattice (c) from figure 4).[8]

The value n is attached to actions that are deontically unsaturated (have no deontic value at all, are plainly neutral). $\top$ is attached to actions that are deontically over-saturated (have obligatory and forbidden components). Each of them is neither "purely" obligatory nor "purely" forbidden, and in that sense is neutral.

Formally, the operator "$\sqcap$" is interpreted as the infimum in the structure (see table 3). Moreover, negation of $\top$ is $\top$ and negation of n is n (see table 4).

[7] One can easily check that in the appropriate trivalent logic

$$(o \sqcap o) \sqcap f \neq o \sqcap (o \sqcap f)$$

[8] The construction in the context of deontic logic occurs also in [4, 12].

$\sqcap$	f	n	o	$\top$
f	f	f	$\top$	$\top$
n	f	n	o	$\top$
o	$\top$	o	o	$\top$
$\top$	$\top$	$\top$	$\top$	$\top$

Table 3.

α	$\overline{\alpha}$
f	o
n	n
o	f
$\top$	$\top$

Table 4.

α	$\mathsf{O}(\alpha)$	$\mathsf{F}(\alpha)$	$\mathsf{N}(\alpha)$
f	0	1	0
n	0	0	1
$\top$	0	0	1
o	1	0	0

Table 5.

Table 5 establishes a correspondence between deontic values $f, n, \top$ and o and deontic operators (where 1 and 0 stand for "true" and "false", respectively). Intuitively, values n and $\top$ are both treated as *neutral*—we can see that $\mathsf{N}(\alpha)$ is true when α is n or $\top$. Thus, in a sense, the system remains trivalent, though formally there are four values that can be attached to actions.

In [12, Section 4.3] there has been introduced an axiomatisation of the logic that is sound and complete with respect to the matrices 3–5.

5.5. Formula "D" and the possibility of conflicts between obligations

A counterpart of the modal logic D axiom

$$\neg(\mathsf{O}(\alpha) \wedge \mathsf{O}(\overline{\alpha})) \tag{2}$$

is valid in the system described above. Taking that into consideration one may wonder whether formula (2) does not "rule out the possibility of (unresolved) conflicts between obligations" and in general one may find "hard to see how it can be claimed that the logic accommodate such conflicts".

Continuing the explanation conducted in Section 5.3 we add that formula (2) only says that one and the same atomic action kind α cannot be subject to conflicting obligations.

In our proposal normative conflicts can emerge *only* as a result of aggregating two or more action kinds (that are subjects to norms). For instance we may have two action kinds α and β and assume that they are subjects of two norms $\mathsf{O}(\alpha)$ and $\mathsf{O}(\overline{\beta})$. As long as there is a way of doing α without doing β, there is no conflict. Otherwise the deontic value of any action token that is classified as $\alpha \sqcap \beta$ (i.e., it is at the same time classified as α and β) should be of an *all-things-considered* nature. In our system, the formula below is valid:

$$\mathsf{O}(\alpha) \wedge \mathsf{O}(\overline{\beta}) \to \mathsf{N}(\alpha \sqcap \beta)$$

It classifies $\alpha \sqcap \beta$ as neutral, providing $\mathsf{O}(\alpha)$ and $\mathsf{O}(\overline{\beta})$.

Below we provide a few facts about aggregation of action kinds in our logic.

Firstly, the formula

$$\neg(\mathsf{O}(\alpha \sqcap \beta) \wedge \mathsf{O}(\overline{\alpha \sqcap \beta}))$$

is not a well-formed formula of our logic. It is so because a complement of action kinds aggregation, e.g., "$\overline{\alpha \sqcap \beta}$", is not allowed in the language.

Moreover, for any compound description of action $\alpha \sqcap \beta$ the formulas below are valid but their inverses (i.e., implications from right to left) are not:

$$\mathsf{O}(\alpha) \wedge \mathsf{O}(\beta) \to \mathsf{O}(\alpha \sqcap \beta)$$
$$\mathsf{F}(\alpha) \wedge \mathsf{F}(\beta) \to \mathsf{F}(\alpha \sqcap \beta)$$
$$\mathsf{N}(\alpha) \wedge \mathsf{N}(\beta) \to \mathsf{N}(\alpha \sqcap \beta)$$

Thus from the fact that $\alpha \sqcap \beta$ is obligatory (forbidden or neutral), it does not follow that each component of this aggregation should be obligatory (forbidden or neutral).

And last but not least, it is falsifiable that if some action type is obligatory (forbidden or neutral), then it is such in aggregation with any other action type:

$$\mathsf{O}(\alpha) \to \mathsf{O}(\alpha \sqcap \beta)$$
$$\mathsf{F}(\alpha) \to \mathsf{F}(\alpha \sqcap \beta)$$
$$\mathsf{N}(\alpha) \to \mathsf{N}(\alpha \sqcap \beta)$$

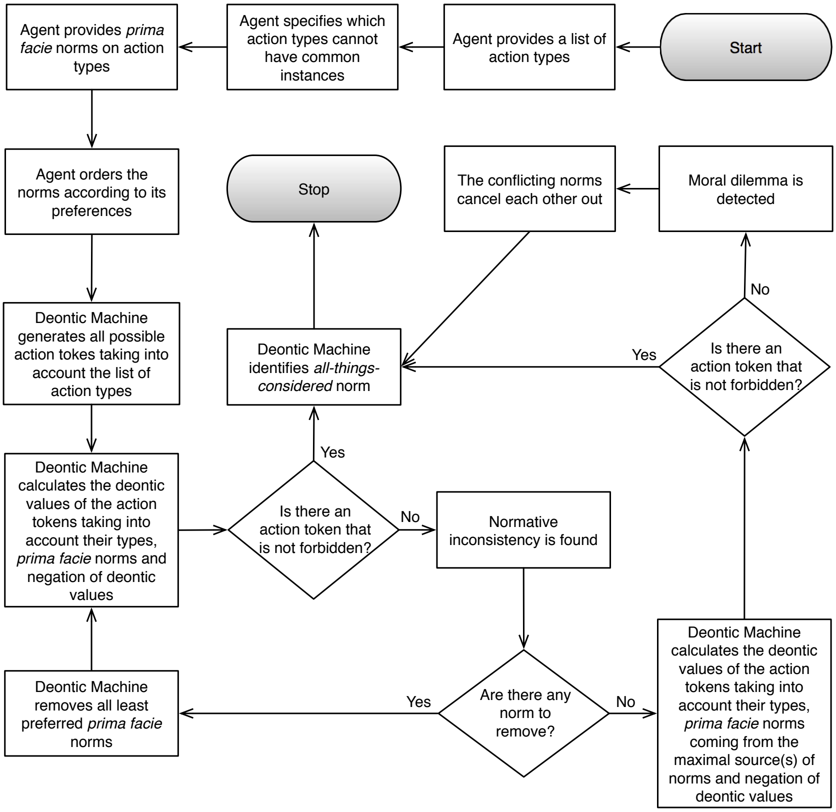
Start
Agent provides a list of action types
Agent specifies which action types cannot have common instances
Agent provides *prima facie* norms on action types
Agent orders the norms according to its preferences
Deontic Machine generates all possible action tokes taking into account the list of action types
Deontic Machine calculates the deontic values of the action tokens taking into account their types, *prima facie* norms and negation of deontic values
Is there an action token that is not forbidden?
Yes
No
Deontic Machine identifies *all-things-considered* norm
Stop
Normative inconsistency is found
Are there any norm to remove?
Yes
No
Deontic Machine removes all least preferred *prima facie* norms
Deontic Machine calculates the deontic values of the action tokens taking into account their types, *prima facie* norms coming from the maximal source(s) of norms and negation of deontic values
Is there an action token that is not forbidden?
Yes
No
Moral dilemma is detected
The conflicting norms cancel each other out

Figure 6. A flowchart diagram representing an algorithm constituting the Deontic Machine.

6. Deontic Machine

In figure 6 there is a flowchart representing in an informal way an algorithm constituting our Deontic Machine. The sequential steps of the program execution are represented as rectangles or diamonds and their order is organized by means of arrows.

The first four steps require data inputted by the user: action types that apply to a situation, *prima facie* duties and an order of the norm sources. After the user-data is collected the deontic machine looks for the best action token to be carried out.

Let us start with the formal description of Example 2.1.

```
%ONTOLOGY: ACTION TYPES

%action types
action_type(save_passengers).
action_type(save_pedestrians).

%colisions
impossible_together([save_passengers,save_pedestrians]).

%DEONTOLOGY: DESCRIPTION OF A NORMATIVE SYSTEM

%the content of norms
norm(passengers,save_passengers,obligatory).
norm(pedestrians,save_pedestrians,obligatory).

%preference relation on norms
prefer(passengers,pedestrians).
```

Running the deontic machine program gives the following results.[9]

```
2 ?- machine.
Discernible tokens:
a1: [save_passengers,n(save_pedestrians)]
a2: [n(save_passengers),save_pedestrians]
a3: [n(save_passengers),n(save_pedestrians)]
System with full set of norms is inconsistent.
System is consistent after removing some norms.
Active norms:
norm(passengers,save_passengers,obligatory)
Removed norms:
norm(pedestrians,save_pedestrians,obligatory)
Obligatory set of tokens:
[a1]
true.
```

Let us now pass to Example 2.2 which differs from the previous example in the fact that there is no preferences between norms. Formally it can be described in the following way.

[9] Both examples presented in this paper and the Prolog code of Deontic Machine are avaliable here: https://kpi.kul.pl/selfdrivingcar. One may also compare it with our Deontic Machine that has implemanted three multi-valued deontic logics: https://kpi.kul.pl/deonticmachine.

```
%ONTOLOGY: ACTION TYPES

%action types
action_type(save_old_man).
action_type(save_child).

%colisions
impossible_together([save_old_man,save_child]).

%DEONTOLOGY: DESCRIPTION OF A NORMATIVE SYSTEM

%the content of norms
norm(pedestrians,save_old_man,obligatory).
norm(pedestrians,save_child,obligatory).
```

Now, running the deontic machine program gives the following results.

```
2 ?- machine.
Discernible tokens:
a1: [save_old_man,n(save_child)]
a2: [n(save_old_man),save_child]
a3: [n(save_old_man),n(save_child)]
System with full set of norms is inconsistent.
System cannot be repaired
Active norms:
norm(pedestrians,save_old_man,obligatory)
norm(pedestrians,save_child,obligatory)
Removed norms:
no removed norms
Best choices indicated by logic:
[a1,a2]
true.
```

7. Conclusions

Designing computer systems that control self-driving cars in accordance with transparent ethical principles is an important challenge today. We have presented a solution based on deontic logic. Ross's conception of *prima facie* and *all-things-considered* duties fits well with the discussion about the ethical issues concerning self-driving cars. A preference order on norms and multi-valued logic is useful for reasoning in this context.

Using these ideas we have built our Deontic Machine to show how they can work in practice.

The system is open to different ethical preferences. They can be encoded by specifying a set of norms and a preference relation on those norms from that set.

Our Deontic Machine always finds an answer regardless of whether or not the input data is a consistent set of norms. Moreover, it has implemented a multi-valued logic that deals with the inconsistency of norms in such a way that it liberates a self-driving car from responsibility in a conflicting situation in which the conflict cannot be solved by preferences.

Acknowledgements. This research was supported by the National Science Centre of Poland (grant UMO-2015/17/B/HS1/02569). We would like to thank Guido Governatori for his valuable remarks after our presentation during DEON 2016 and Michael Musielewicz for all discussions about autonomous cars. We would also like to thank anonymous reviewers for their valuable remarks. Some of them that we could not address adequately in the present paper will be an inspiration for future work.

References

[1] Belnap, Nuel, "A useful four-valued logic", pages 8–37 in J. M. Dunn and G. Epstein (eds.), *Modern Uses of Multiple-Valued Logic*, Springer, 1977.

[2] Boella, Guido, and Leendert W.N. van der Torre, "A game-theoretic approach to normative multi-agent systems", in *Normative Multi-agent Systems*, 2007. URL http://icr.uni.lu/leonvandertorre/papers/normas07a.pdf

[3] Contissa, Giuseppe , Francesca Lagioia and Giovanni Sartor, "The ethical knob: Ethically-customisable automated vehicles and the law", *Artif. Intell. Law* 25, 3 (2017): 365–378. DOI: 10.1007/s10506-017-9211-z

[4] Craven, Robert, "Policies, norms and actions: Groundwork for a framework", Technical report, Imperial College, Department of Computing, 3 2011.

[5] De Haan, Jurriaan, "The definition of moral dilemmas: A logical problem", *Ethical Theory and Moral Practice* 4, 3 (2001): 267–284. DOI: 10.1007/10.1023/A:1011895415846

[6] Goble, Lou, “A logic for deontic dilemmas”, *Journal of Applied Logic* 3, 3–4 (2005): 461–483. DOI: 10.1007/10.1023/10.1016/j.jal.2005.04.004

[7] Hansen, Jörg, “Deontic logics for prioritized imperatives”, *Artif. Intell. Law* 14, 1–2 (2006): 1–34. DOI: 10.1007/s10506-005-5081-x

[8] Hars, Alexander, “Top misconceptions of autonomous cars and self-driving vehicles”, 2016. URL www.inventivio.com/innovationbriefs/2016-09

[9] Holbo, John, “Moral dilemmas and the logic of obligation”, *Americal Philosophical Quarterly* 39, 3 (2002): 259–274.

[10] Horty, John F., “Moral dilemmas and nonmonotonic logic”, *Journal of Philosophical Logic* 23, 1 (1994): 35–65. DOI: 10.1007/BF01417957

[11] Horty, John F., *Agency and Deontic Logic*, Oxford University Press, Oxford, 2001.

[12] Kulicki, Piotr, and Robert Trypuz, “Multivalued logics for conflicting norms”, pages 123–138 in O. Roy, A. Tamminga and M. Willer (eds.), *Deontic Logic and Normative Systems, 13th International Conference, DEON 2016*, 2016.

[13] Sartor, Giovanni, “Normative conflicts in legal reasoning”, *Artif. Intell. Law* 1, 2 (1992): 209–235. DOI: 10.1007/BF00114921

[14] Taylor, Michael, “Self-driving mercedes-benzes will prioritize occupant safety over pedestrians”, October 2016. URL http://blog.caranddriver.com/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/

[15] 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.

[16] van Benthem, Johan, Davide Grossi and Fenrong Liu, “Priority structures in deontic logic”, *Theoria* 80, 2 (2013): 116–152. DOI: 10.1111/theo.12028

[17] Ross, W. D., *The Right and the Good*, Oxford University Press, Oxford, 1930.

Piotr Kulicki and Robert Trypuz
The John Paul II Catholic University of Lublin
Faculty of Philosophy
Al. RacĹÆawickie 14, 20-950 Lublin, Poland
{kulicki,trypuz}@kul.pl