A Default Deontic Logic Analysis of the Trolley Problem in Autonomous Vehicles

Zhaohui Yin

School of the Humanities, China University of Political Science and Law, Beijing 100088, China.

Abstract

Autonomous vehicles, as one of the most common types of machinery that rely on algorithms, have to make value decisions in the "Trolley Problem" situation: a life-saving priority should be given to protecting pedestrians on the road or passengers in the car? When children are involved, how to choose between protecting the children on the road and passengers in the car. From a logical point of view, one important perspective to answer these questions is how intelligent agents like autonomous vehicles can understand the concept of value, such as "good and bad", and then make value judgments and inference. These key issues bear theoretical and practical significance on the future development of autonomous vehicles. Constructing a formal logical framework for the above problems is the beginning for in-depth study. Based on the philosophical premise of value subjectivity and the logic analysis of the autonomous vehicles' NSPR (Norm Set on Power and Responsibility), this article uses the deontic default theory to construct a default model for the intelligent agent and its situation and describes the logical structure of the value judgments made by autonomous vehicles.

Keywords

Deontic Logic; Default Rules; Value Anthropomorphic Subjectivity.

1. Introduction

In today's world, with the booming intelligent technology and rapidly iterating algorithms, technology philosophy is becoming the "first philosophy" in the modern time (Li He, 2020). As artificial intelligence is applied to people's life in an in-depth and extensive manner to take over complicated takes and make various algorithmic decisions, human beings own unprecedently powerful tools for the cognition, utilization, and alteration of the nature and themselves in many important domains. Autonomous driving is considered as one of the effective and promising domains in today's artificial intelligence development, and is regarded as the "national strategy" of artificial intelligence development by the nations such as the United States of America, European Union, and China in sequence. The popularization of autonomous driving will not only abstract human beings from machinery labor, but also obviously improve overall safety on the road (*The Global Status Report on Road Safety* 2018 by the WHO indicates that there are about 1.35 million people die in road accidents, among which 90% have human factors.). Thus, the problems related with autonomous driving, such as technology, law, safety, and ethics, are extensively focused on and talked about. Among these problems, the trolley problem in autonomous vehicles is typically an ethical problem with conflicts in value and norm. Because conflicts in norm often lead to complicated scenes, the reasoning ability in such situations is considered as an important "machine intelligence representation". Therefore, the research by attempting constructing the logical framework of scenes with conflicts in norm to depict the corresponding reasoning behavior is helpful in the development of artificial intelligence and other ethical issues.

2. Trolley Problem in Autonomous Vehicles: A Problem of Using Algorithm to Distribute Life

When the "trolley problem", a thought experiment of traditional philosophers, becomes people's daily practices, and the independent-decision-making intelligent machine takes over the decision-making entity from natural persons controlling electronic vehicles, should the safety of passengers in the

vehicle be considered as the supreme behavior principle? Or can we forgive the intelligent machines for following the ethical instinct to comply with the utility theory of "saving the many"? The "Trolley Problem in Autonomous Vehicles", or the "Tunnel Problem" of autonomous vehicles (Gogoll, 2017) is defined as what logic autonomous vehicles follow to make "good" decisions or decisions based on correct value, under the scene with conflicts in norm, that the safety of passengers and pedestrians cannot be both guaranteed. Academic scholars hold different vies over the relevance of the classical trolley problem to the problem in autonomous vehicles. For example, Nyholm (2016) believes that the two problems can hardly be the same, because there are differences in aspects such as cognitive context, information acquisition, and ethical principle applications. However, this article proposes the belief that though there are differences in the two types of trolley problems, especially in the subject (natural person with full autonomy vs. intelligent machine with relative autonomy), the logical constructs of the two are identical. In the two types of problems, the dilemma is both caused by the conflicts in value or ethical norm, and this is the object for logic analysis in this article. Thus, the difference between the two trolley problems will not actually affect the form depiction of the common logic construction.

The trolley problem in autonomous vehicles is actually a practical ethical problem using algorithms to distribute life. The trolley problem in autonomous vehicles is different from the "classical trolley problem" in the ideal scene, regarding the ethical choice as the core, and also different form the traditional road accidents in the real situation, regarding the after-incident responsibility distribution. The uniqueness of the trolley problem in autonomous is that the distribution for human safety in a certain scene is strictly conducted by a certain pre-configurable algorithm based on a certain value principle, through intelligent machines (Tasioulas, 2019). Neither the identities of traffic participants nor the involvedness of traffic participants will have direct influence on the trolly problem in autonomous vehicles. For example, according to certain researches, the identities of "safe bystander" and "potential victim" have great influence on the choice between the utility theory and the deontology as the value principle of autonomous vehicles. For example, according to the latest research results of Bergmann (2018), when the experiment participants are faced with a scene of conflicts in norm, with one passenger against five pedestrians, 76% of the experiment participants (safe bystanders) choose the value principle of sacrificing the passenger but saving the pedestrians for autonomous vehicles. This complies with the utility theory principle and is consistent with the results of the classical trolley problem experiment carried in Harvard University in 2003 (when faced with the trolley problem, 89% people choose the principle of the utility theory). Based on the preceding results, if the identity is "safe bystander" in a trolley problem, the "ethical instinct" of utility theory of "saving the many" is accepted by most people. However, the results change completely when the experiment participants' identity is changed from the safe bystander to the potential victim. In 2006, the Harvard University professor Bonnefon (2016) *et al* have published a paper on *Nature*. In their research on 2.3 million people in the globe, there are only a few showing the intent of purchasing autonomous vehicles with the utility theory principle preconfigured. Thus, if people become "potential victims", most people still have difficulties in accepting being in an autonomous vehicle using the utility theory as the value principle to handle the scene of conflicts in norm.

3. Logic: From Classical Deontic Logic to Default Deontic Logic

Currently, the researches on the trolley problem in autonomous vehicles focus on a positivism path of a bottom-up approach. This path origins from the prosperity of machine learning algorithms based on big data, and the core is to make intelligent machines to "learn" some key specifications from massive samples, and to make algorithmic decisions accordingly. For example, Leben (2017) has noted in the article *A Rawlsian algorithm for autonomous vehicles*, that the Max Min Principle described in *the Theory of Justice* by Rawls should be used as the ethical principle for the trolley problem in autonomous vehicles, and he has designed an "accident algorithm" accordingly. In the algorithm scheme of Leben, a cartesian product is produced by the utility function of the accident subjects (passenger and pedestrians) and the candidate behaviors (going straight and turning the

vehicle) when the trolley problem in autonomous vehicles occurs, in the scene of conflicts in norm. The lowest benefit set for "surviving probability" of the accident is located from the mapped data set of the cartesian product, and a cyclic exhaustion is implemented to make the operation decision with the minimum benefits maximized. The autonomous vehicle performs the driving behavior based on this decision. Another example, Grau (2006) has designed an algorithm based on the "principle of minimizing the overall damage" of the utility theory. However, Coca-Vila (2017) does not accept the utility theory as the algorithmic principle, and believes that deontology is more proper to be the algorithmic principle of autonomous vehicles.

Indeed, the inductive algorithm with positivism of the bottom-up approach, which depends on massive data "feeding" on technology itself, has great advantages in certain fields. However, this machine learning algorithm generated and evolving highly depending on big data has a fundamental dependency on the data quality and quantity, underlying the efficiency and effectiveness of such a data-depending machine learning algorithm, but this dependency brings the difficulty. For example, sufficient fixed, limited, and manually tagged "representative" data are is to effectively train the relevant machine learning algorithm, so the output correctness rate reaches a certain standard (Chen Xiaoping, 2020). If the closed requirement for the scene system is not satisfied, the effectiveness of the machine learning algorithm application cannot be guaranteed. In addition, Hinton, Bengio, and LeCun (2015), the three most eminent machine learning scientists in the world, have clearly pointed out that the fundamental defect of machine learning algorithm is the lack of complicated deductive reasoning ability. The difficulty as well as the core of the complicated deductive reasoning is how to provide effective ability of commonsense reasoning, and causal reasoning. For example, because of the complicated real road traffic for the autonomous vehicles, the autonomous algorithm needs to have the deductive reasoning ability based in complicated scenes described by incomplete and defeated data. However, the current machine learning algorithms can hardly have such ability. Moreover, the biggest challenge of machine learning algorithms is the algorithmic "black box" caused by the algorithmic opacity (Ding Xiaodong, 2020). Such opacity is largely determined by the technological characteristics of deep learning algorithms, but it also causes confusion to people's clear understanding on the algorithm operations, and hampers the effective regulation of rules in the algorithm in a certain degree.

In the view of modern logic, the top-down deductive reasoning logic constructing approach is also available for the research on the trolley problem in autonomous vehicles. Using the technological method of deontic logic to transfer the regulated behaviors of intelligent machines into a formalized logic calculation is the presentative solution of this approach. This solution origins from the deep integration of the deontic logic development in the recent few decades with the artificial intelligence and computing science, and the deep integration is considered as one of the most influential interdisciplinary fields to various types of behavior researches of law and ethics (Meyden, 2012). The deontic logic is also often called the normative logic. The form research of norm and its reasoning is not only the initial intention of creating the deontic logic, but also the primary method of using the deontic logic to study artificial intelligence problems. Therefore, though the deductive algorithm such as the deontic logic has far lower operation efficiency than that of inductive algorithms of machine learning, the deductive algorithm has the advantage that multiple and multi-layer logic systems can be constructed and be put into collaborative operation to enable the corresponding algorithms to have much capability as possible of processing complicated problems (Except the solutions in the preceding two paths, Allen (2008), a well-known artificial intelligence philosopher, has proposed that philosophers and computing engineers should co-design three paths for Automated Moral Agents (AMAs): up-to-bottom, bottom-to-up, and up-bottom-mixed paths.). This is mutually and effectively complementary with the machine learning algorithm. Though we do not believe that the deontic logic will surely provide complicated deductive reasoning ability or be complementary with intelligent machine algorithms, simplifying the complicated determination process of value or ethics into the precise reasoning based on logic calculus is still a vital approach of completing the deductive reasoning ability in autonomous vehicles.

Many classical deontic logic systems, nevertheless, have perfect form characteristics, that is, pretty good reliability and completed results (Yu Junwei, 2005). However, the deontic logic depicts real behavior norm, and therefore we must focus on not only the system characteristics themselves, but also the depiction effect of the behavior norm, that is, the matching of deontic depiction of norm and the intuitive understanding. On this matter, there is still some deficiency of the classical deontology theoretical basis, producing various types of "deontic paradoxes", greatly hampering the development of the classical deontic logic. The classical deontic logic is a monotonic reasoning lack of "fault tolerance" or "defeat capability", making it insufficiently capable of processing the preceding problems in daily reasoning of norm reasoning.

Default deontic logic is an improvement of the syntax and semantics of the system with the preceding problems for the classical deontic logic. The default deontic logic is the deontic logic with the default rules included. The default rules are a certain linguistic expression which can be intuitively construed in the following way: if there is no sufficient evidence indicating that a proposition is false, then it is true. If a counter-example occurs, the proposition and related conclusions are revoked. The default system is firstly proposed by Reiter (1980), an artificial intelligence expert, to study non-monotonic reasoning. The deontic default rules process conflicts in norm or duty, making deontic system to tolerate conflicts in a certain degree. In recent years, the fast development of artificial intelligence and cognitive science also further completes the default deontic logic. Therefore, the default deontic logic becomes a vital choice for studying complicated norm reasoning.

4. The Default Deontology Scheme by Horty

In this section, the extensively influential default deontology theory by Horty, the well-known logician, is introduced. With this theory, a default semantic model with an intelligent entity and scene is established, to analyze the value determination and logic structure of autonomous vehicles in the "trolley problem", and extend and reveal the value characteristics of the default system.

Horty (1994, 2001, 2003, 2014) discards the classical deontic semantic theory, and proposes a semantic scheme based on the default theory. The core of the semantic scheme by Horty is that in a specific deontic scene, the default theory can be used to determine the true value of a deontic sentence after the scene and norm are described (Horty, 2014). Horty defines the condition of true value for the deontic sentence:

$\mathcal{O}\alpha$ is true for the norm system $(A, <)$, if and only if α greatly satisfies the norm subset of the top priority.

The key of the scheme is how to understand the norm and the norm system. Horty regards the norm system as an imperative sentence set with partial ordering relation. For example, $(A, <)$ is used to indicate a norm system, with $(A, <)$ indicating the norm set and $<$ indicating the partial ordering relation representing the priority relation of different norms. It is worth noting that Horty uses the default implication equation to present the form of norm. The default implication formula is a type of implication which allows exceptions. For example, $\beta \Rightarrow \alpha$ can be interpreted as the following: If Tweedy is a bird, Tweedy can fly. Obviously, this example has exception (if Tweedy is a penguin).

Horty further defines the default theory $\Delta = (W, D, <)$. The default theory is a triple, with W indicating a finite descriptive sentence set of narrative facts, D indicating a finite default implementation formula set and $<$ indicating the partial ordering relation of D. For example, $\phi < \psi$ indicates that ϕ has priority over ψ. If $r = \delta \Rightarrow \sigma$ is defined as an arbitrary value norm, its antecedent is defined as $Pre(r)$, which is indicated by δ in this value norm, and its consequent is defined as $Con(r)$, which is defined as σ in this value norm. Particularly, the following equation is workable for an implication formula: $Con(A) = \{ Con(\psi) \mid \psi \in A \}$. A scenario S based on the default deontology theory Δ is the subset of D, and the following three conditions must be met for the restriction scenario S: triggered, not conflicted, and undefeated. Details are as follows:

(1) "Triggered" indicates that if a default rule is used for reasoning, the default rule must be possibly violated. In scenario S, the triggered default implication formulas belong to the following set:

$$Triggered_{\{W,D\}}(S) = \{\phi \in D: W \cup Con(S) \vdash Pre(\phi)\}.$$

This indicates that in a certain situation, if the antecedent of a rule can be inferred from the description of the situation and the conclusion of scenario S, the rule can be triggered.

(2) "Not conflicted" indicates that the default rules used in reasoning must be consistent. In scenario S, the conflicted default implication formulas form the following set:

$$Conflicted_{\{W,D\}}(S) = \{\phi \in D: W \cup Con(S) \vdash \neg Con(\phi)\}.$$

This indicates that in a certain situation, if the denial of an antecedent for a rule can be inferred from the description of this situation and the conclusion of scenario S, the rule has contradiction and therefore is conflicted. Whereas, the rule is not conflicted.

(3) "Undefeated" indicates that default rules used in reasoning will not be defeated by other default rules with higher priority.

$$\begin{aligned} Defeated_{\{W,D,<\}}(S) = \quad & \{\phi \in D: \exists \psi_0 \dots \psi_n \in Triggered_{\{W,D\}}(S) \\ & \text{Therefore, any } i \in [0, \dots, n], \psi_i < \phi \\ & \& W \cup Con(\{\psi_0, \dots, \psi_n\}) \vdash \neg Con(\phi)\}. \end{aligned}$$

This indicates that if some rules are triggered in scenario S, the priority of the rules is higher than that of rule ϕ. In this case, if the antecedent of rule ϕ can be inferred through the description of this situation and the consequent of these rules, rule ϕ is defeated. Whereas, the rule is not undefeated.

In addition, the binding set in scenario S is as follows:

$$\begin{aligned} Binding_{\{W,D,<\}}(S) = \{\phi \in D: \quad & \phi \in Triggered_{\{W,D\}}(S); \\ & \phi \notin Conflicted_{\{W,D\}}(S); \\ & \phi \notin Defeated_{\{W,D,<\}}(S)\}. \end{aligned}$$

The preceding set indicates that if a rule is triggered, is not conflicted in scenario S, and is undefeated, the binding takes effect for the rule in scenario S.

In this way, a proper scenario S with the preceding three conditions met, needs to comply with the formula $S = Binding_{\{W,D,<\}}(S)$. Such a scenario S is called a default extension for the default theory $\Delta = (W, D, <)$. The logic consequence of the default extension, intuitively speaking, is the proposition "complying with" the norm system.

5. The Default Deontic Model for the Trolley Problem in Autonomous Vehicles

We put forward a philosophical assumption: intelligent entities, including autonomous vehicles, have a value subjectivity, which is called Value Quasi-subjectivity of Intelligent Agents. The quasi-subjectivity is indicated by the following: human beings specify the independent value judgement function in certain scenes on various types of intelligent entities, and therefore the intelligent entities can play the role of the value entity in the related scenes.

Based on the preceding description, if the intelligent entity is considered as an intelligent machine with value quasi-subjectivity, an arbitrary intelligent entity, according to Horty's default scheme, can uniquely map to a group of ordered pairs $(D, <)$. D indicates the set of norms of rights and liabilities, and $<$ indicates the value position of the intelligent entity, which is represented by a partial ordering relation. If any group of ordered pairs is presumably able to uniquely mark an intelligent entity, an arbitrary situation of conflicts in norm is a triple $(W, D, <)$, where W is the narrative sentence set describing the situation of conflicts, and $(D, <)$ is the intelligent entity.

The trolley problem in autonomous vehicles is a specific situation of conflicts in norm, and an arbitrary trolley problem in autonomous vehicles can be defined as a tripe $\Delta 1 = (W, D, <)$, where:

$\boldsymbol{W}$ is the narrative sentence set describing the trolley problem scene in autonomous vehicles. For example:

1) $\phi 1$: there are 5 pedestrians in front on the road.

2) $\phi 2$: there is 1 passenger in the vehicle.

3) $\phi 3$: if the autonomous vehicle continues the predefined driving path, it will kill 5 pedestrians in a crash, but the 1 passenger in the vehicle will live.

4) $\phi 4$ if the autonomous vehicle takes a sharp turn, it will crash into the building on the side of the road and kill the 1 passenger in the vehicle, but the 5 pedestrians on the road will live.

5) $\phi 5$: the autonomous vehicle has only two options, continuing the predefined driving path, and taking a sharp turn.

$\boldsymbol{D}$ is the set of norm of rights and liabilities, including all its value norms. For example:

1) $D1$: avoid harm the 5 pedestrians on the road.

2) $D2$: avoid harm the 1 passenger in the vehicle.

3) $D3$: if there are kids in and only in pedestrians, avoid harm these kids.

$<$ is the partial ordering relation of the autonomous vehicle. The partial ordering relation indicates that the value positions represent the orientation of different values. For example:

1) If $D1 < D2$, then the corresponding value orientation $V1$ is: "trying to protect most people".

2) If $D3 < D2$, then the corresponding value orientation $V2$ is: "trying to protect children".

3) If $D2 < D1$, then the corresponding value orientation $V3$ is: "trying to protect the passenger in the vehicle".

Before the preceding value judgement for the intelligent machine in the trolley problem in autonomous vehicle $\Delta 1$, a pair of basic value operators need to be defined. We use the logic operator $\mathcal{G}$ to indicate the "good" positive value, and operator $\mathcal{B}$ to indicate the "bad" negative value. For an arbitrary proposition variable p, $\mathcal{G}p$ and $\mathcal{B}p$ are called "p is good" and "p is bad", respectively.

Under an arbitrary scene of conflicts in norm $(W, D, <)$:

I) "p is good" can be recorded as $(W, D, <) \vDash \mathcal{G}p$.

II) "p is bad" can be recorded as $(W, D, <) \vDash \mathcal{B}p$.

According to the deontology default semantics, whether p is good or bad is determined by the triggered value norm in the highest position. For an arbitrary scene with conflicts in norm $(W, D, <)$ and value norm $r \in D$, if $W \vdash Pre(r)$, value is triggered for r in $(W, D, <)$; in the condition of the value triggered for r, for the value norm $r' \in D$ of an arbitrary value triggered, if $r' \in Binding_{\{W,D,<\}}(S)$, then $r < r'$, and the value maximum if triggered for r in $(W, D, <)$.

Thus, we can define whether any proposition is good or bad in a scene of arbitrary conflicts in norm, that is, the form definition of value judgement.

Definition 1: assume that $(W, D, <)$ is an arbitrary scene of conflicts in norm, and ϕ is an arbitrary proposition logic formula,

I) $(W, D, <) \vDash \mathcal{G}\phi$, if and only if $r \in D$ exists and value maximum is triggered for r, and $W \cup \{Con(r)\} \vdash \phi$.

II) $(W, D, <) \vDash \mathcal{B}\phi$, if not and only if not $(W, D, <) \vDash \mathcal{G}\phi$.

Actually whether an arbitrary proposition in a scene of conflicts in norm is "good" or "bad" can be interpreted in two ways: an arbitrary proposition is good if it complies with the value norm of the highest position, or an arbitrary proposition is good if it complies with the value norm system. Definition 1 in the text has depicted the first interpretation. However, according to the work of Horty in the preceding section, the following definition can be introduced to depict the second interpretation.

Definition 1': assume that $(W, D, <)$ is an arbitrary scene of conflicts in norm, and ϕ is an arbitrary proposition logic formula,

I) $(W, D, <) \vDash \mathcal{G}\phi$, if and only if the default expansion S of $(W, D, <)$ causes $W \cup S \vdash p$.

II) $(W, D, <) \vDash \mathcal{B}\phi$, if not and only if not $(W, D, <) \vDash \mathcal{G}\phi$.

Detailed comparison between definition 1 and definition 1' is not conducted here, and this is only a reminder for you that definition 1' is a value judgement of "good" and "bad" obtained from a more subtle examining the relationship between the norm system and the proposition. This allows us to introduce the value comparison operators of "better" and "worse" in the following discussion. Definition 1 is of ease and sufficiency in use for scenes of common conflicts in norm.

We select two possible situations to analyze the value judgement of intelligent machines in the trolley problem in autonomous vehicles $\Delta 1$.

Situation I: assume that the 5 pedestrians on the road are all adults. According to $\Delta 1$, value will be triggered for $D1$ and $D2$, and value not be triggered for $D3$ due to the precondition "kids in pedestrians". That is, $Triggered_{\{W,D\}}(S) = \{D1, D2\}$. If the partial ordering relation of $\Delta 1$ is $D2 < D1$, then $Defeated_{\{W,D,<\}}(S) = \{D1\}$, and consequently $Binding_{\{W,D,<\}}(S) = \{D2\}$. Therefore, the value maximum is triggered for $D2$. Then the following can be concluded: $W \cup \{Con(D2)\} \vdash Con(D2)$, $W \cup \{Con(D2)\} \vdash \neg Pre(\phi 4)$, $W \cup \{Con(D2)\} \vdash Pre(\phi 3)$. Based on definition 1, $(W, D, <) \vDash \mathcal{G}(Pre(\phi 3))$.

From the preceding conclusions, the intelligent machine of autonomous vehicle following the value orientation $V3$ "trying to protect the passenger in the vehicle" will make the following value judgement when encountering the situation that "all 5 pedestrians on the road are all adults": "'continuing the predefined driving path' is good." The autonomous vehicle in the trolley problem $\Delta 1$ will make the decision of continuing the predefined driving path based on this value judgement.

Situation II: assume that there are and only are kids in pedestrians. According to $\Delta 1$, the precondition of $D3$ is true. If the assumed partial ordering relation in $\Delta 1$ is $D3 < D2$, the value, especially the value maximum, is triggered for $D3$. That is $Triggered_{\{W,D\}}(S) = \{D3\}$. Then the following can be concluded: $W \cup \{Con(D3)\} \vdash Con(D3)$. According to definition 1, the following can be concluded in the same way: $(W, D, <) \vDash \mathcal{G}(Pre(\phi 4))$.

From the preceding conclusions, the intelligent machine of autonomous vehicle following the value orientation $V2$ "trying to protect kids" will make the following value judgement when encountering the situation that "there are and only are kids": "'taking a sharp turn' is good." Autonomous vehicles in the trolley problem $\Delta 1$ will make the decision of taking a sharp turn based on this value judgement.

The following theorem can be concluded from definition 1:

Disjunctive theorem: assume that $(W, D, <)$ is an arbitrary scene of conflicts in norm, and ϕ and ψ are arbitrary proposition logic formulas. In this situation, if $(W, D, <) \vDash \mathcal{G}\phi$ or $(W, D, <) \vDash \mathcal{G}\psi$, then $(W, D, <) \vDash \mathcal{G}(\phi \vee \psi)$.

The disjunctive theorem stands depending on the closure of the logic consequence "$\vDash$" in definition 1 and the introduction of the disjunctive rule to the logic consequence. According to the disjunctive theorem, in a certain scene of conflicts in norm, if at least one of two arbitrary propositions is good, the disjunction of two propositions are good. For example, in the trolley problem in autonomous vehicles $\Delta 1$, if we have now clear information about a specific situation of the value norm in the set of norm of rights and liabilities for the autonomous vehicle, but know the following value judgement: "'the autonomous vehicle continuing the predefined path' is good, or 'the autonomous vehicle taking a sharp turn' is good." Then, the following can be concluded according to the disjunctive theorem: "'the autonomous vehicle continuing the predefined driving path or taking a sharp turn' is good."

However, the following proposition does not stand according to either definition 1 or definition 1': if $(W, D, <) \vDash \mathcal{G}\phi$ and $(W, D, <) \vDash \mathcal{G}\psi$, then $(W, D, <) \vDash \mathcal{G}(\phi \wedge \psi)$. The reason that this proposition does not stand is that there are quantifiers limiting the norm of great position in definition 1 and the default extension of definition 2. That is, in a certain scene of conflicts in norm, if one of two arbitrary propositions is good in a certain scene of conflicts in norm, then it cannot be inferred that both of the two propositions are good. For example, psychologists propose that both moderate drinking and taking a ride can ease the dismay of a person. Thus, it can be inferred for a person with dismay,

moderate drinking is good, and taking a ride is also good. However, it cannot be concluded that the proposition of "drinking and taking a ride" is good.

Apart from the preceding analysis, there are other value operators distributed in a scattered or consecutive manner between the two-value operators of "good and bad" or "true or false" in actual value judgement. For example, people often use the value operators of comparative degrees such as "better" and "worse" for value comparison. For example, "'continuing the predefined driving path' is better". Therefore, there can also be comparability in the value judgement of autonomous vehicles. Through a more subtle depiction of the partial ordering relation in the set of norm of rights and liabilities for the intelligent entity, we give definitions of "better" and "worse" for different value norm weight to present the comparability of value.

Specifically, the function from the set of norms of rights and liabilities to the set of real numbers $L: D \to Q$ to indicate the weight function of value norm. In this function, Q indicates the set of real numbers. Through the weight function L, we can draw the partial ordering relation $<$ on D from the following: for arbitrary $r_1, r_2 \in D$, $r_1 < r_2$ if and only if $L(r_1)\ is\ equal\ to\ or\ greater\ than\ L(r_2)$. The partial ordering relation $<$ is the causal relationship for L. For an arbitrary set $A \subseteq D$, $L[A]$ is the sum of weight for all value norms in A, that is $\Sigma\{L(r): r \in A\}$. Then, "better" ($\prec$) can be defined as follows:

Definition 2: assume that (W, D, L) is an arbitrary scene of conflicts in norm, and $\phi\ and\ \psi$ are arbitrary proposition logic formulas.

$(W, D, L) \vDash \psi \prec \phi$ if and only if:

1) $(W, D, L) \vDash \mathcal{G}\phi$;

2) There is the default extension S for $(W, D, <)$ to form $W \cup Con(S) \vdash \psi$, where $<$ is the causal relationship for L;

3) For an arbitrary default extension S', if $W \cup Con(S') \vdash \phi$, then $L[S]$ is greater than $L[S']$.

With definition 2, the depiction of "better" can make entities to compare value through comparison whether different default extensions are good or bad.

The value judgement of autonomous truck in scene with conflicts in norm is used as example.

Assume that there is an autonomous truck in an artwork warehouse for the transportation of artworks in the warehouse. In a transportation task, the brakes of the truck fail suddenly. In front of the autonomous truck, there is an expensive artwork A on the left, and there is a road block on the right. The cargo B in the autonomous truck has the same price with that of A. If the autonomous truck turns left, A is damaged. If the autonomous truck turns right, the truck and B are damaged. In addition, the autonomous truck is defined to unconditionally protect the artwork from the warehouse, and it should avoid damage of itself to the greatest extent. In this situation, the truck can only turn left or turn right. Then the value norm of the rights and liabilities for the autonomous truck can be assumed as follows:

1) $F1$: keep safe artwork A.

2) $F2$: keep safe artwork B.

3) $F3$: protect the truck itself.

Obviously, the preceding scenes form a scene with conflicts in norm. Each preceding norm has a corresponding price for the protected item. Assume $L(F1) = L(F2) = 100$, and $L(F3) = 1$. Use $\Delta 2 = (W, D, <)$ to indicate the scene with conflicts in norm, where $D = \{F1, F2, F3\}$, and $<$ is the causal relationship of L. Because the autonomous truck can choose to keep safe A or keep safe B and the truck itself, it can be verified that $\{F1\}$ and $\{F2, F3\}$ are both the default extension of $(W, D, <)$. Thus, based on definition 1', turning left and turning right by the autonomous vehicle are both good. However, due to $L(F1) = 100$ and $L(F2) + L(F3) = 101$, based on definition 2, protecting B is better. Therefore, the intelligent entity of the autonomous truck in the scene with conflicts in norm will make the following value judgement: "'turning left' is better." Then, the autonomous truck will turn left.

6. Summary and Forecast

In this article, it is attempted to use the default deontic logic to implement model analysis for the currently focused trolley problem in autonomous vehicles in the artificial intelligence ethical field. From the philosophical assumption that intelligent have the value quasi-subjectivity, how the autonomous vehicle logically maked value judgement to "determine good and bad" in a dilemma.

In this article, the default deontology solution can serve as a reference for the research on the trolley problem in autonomous vehicles, and it is attempted to reveal the positive meaning of modern logical technology on artificial intelligence philosophical problems. For example, from the perspective value philosophy, science is permanent exploration of the boarder of human being's knowledge. Science has value neutrality on the human being entity, but technology is different. Any technology can be regarded as a science with value, and therefore technology itself has no value neutrality. It can be standardized for a certain ethic, and can be regulated in law. However, the rise of machine learning brings great challenge for the two points. For example, the algorithm has challenged the right to be informed of human beings, and threatened the privacy and freedom of individuals, causing discrimination, bias and opacity (To effectively control the risks of artificial intelligence and implementing orderly management of artificial intelligence, multiple countries have released ethical regulations and laws for artificial intelligence. For example, the GDPR (2018) and Trusted Artificial Intelligence Ethical Code (2019) released by the EU, and the New-Generation Artificial Intelligence Development Plan (2018) released by China all propose that the ethical code of "transparency", "explainability" and "accountability" for artificial intelligence should be constructed.). We believe that only when explainability and interpretability are available for the logic base of the algorithm technology, the constructed algorithm system has transparency, and further feasibility of ethical norm and law regulations.

In actual situations, the autonomous vehicle is not isolated, static, and uni-dimensional. It is inevitably an intelligent entity of in a multi-element, diachronic, and multi-dimensional complicated traffic network. In such a traffic network, an arbitrary autonomous vehicle is only one node among many others, which are dynamically connected and make decisions in a collaborative manner. This indicates that to more precisely depict the characteristics of autonomous vehicles, the default deontology framework must be extended in a multi-agent and dynamic way, which will be the questions for further study.

References

[1] Chen Xiaoping, 2020, “Target, Task, and Path for Artificial Intelligence Ethical Construction: Six Topics and Their Basis”, in Philosophical Researches, Volume 9.

[2] Li He, 2020, “Technology From "Agent" to "Substitution" and the Being "Outdated" Human Beings?”, in Social Sciences in China, Volume 10.

[3] Yu Junwei, 2005, “Deontic Logic Study”, in China Social Science Press, 13.

[4] Allen, C., Wallach, W., 2008, Moral Machines: Teaching Robots Right from Wrong, Oxford University Press.

[5] Bonnefon, J. F., Shariff, A., Rahwan, I., 2016, “The Social Dilemma of Autonomous Vehicles”, in Science, 352(6293): 1573-1576.

[6] Bergmann, L. T., Larissa, S., Carmen, M.,2018, “Autonomous Vehicles Require Socio-political Acceptance: An Empirical and Philosophical Perspective on the Problem of Moral Decision Making”, in Frontiers in Behavioral Neuroscience, 12(31): 1-12.

[7] Bergmann, L. T., Larissa, S., Carmen, M.,2018, “Autonomous Vehicles Require Socio-political Acceptance: An Empirical and Philosophical Perspective on the Problem of Moral Decision Making”, in Frontiers in Behavioral Neuroscience, 12(31): 1-12.

[8] Coca-vila, I., 2017, “Self-driving Cars in Dilemmatic Situations:An Approach Based on the Theory of Justification in Criminal Law”, in Criminal Law and Philosophy, 12(1): 1-24.

[9] Gogoll, J., Müller, J. F., 2017, “Autonomous Cars: In Favor of a Mandatory Ethics Setting”. in Science and Engineering Ethics, 23(3): 681-700.

[10] Brogan, A. P., 1919, “The Fundamental Value Universal”, Journal of Philosophy, in Psychology and Scientific Methods, vol.16: 96-104.

[11] Gogoll, J., Müller, J. F., 2017, “Autonomous Cars: In Favor of a Mandatory Ethics Setting”. in Science and Engineering Ethics, 23(3): 681-700.

[12] Horty, John F., 1994, “Moral dilemmas and nonmonotonic logic”, in Journal of Philosophical Logic, 23(1): 35-65.

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

[14] Horty, John F., 2003, “Reasoning with moral conflicts”, in Noûs, 37 (4): 557-605.

[15] Horty, John F., 2014, “Deontic Modals: Why Abandon the Classical Semantics?”, in Pacific Philosophical Quarterly, 95(4): 424-460.

[16] Leben, D., 2017, “A Rawlsian Algorithm for Autonomous Vehicles”, in Ethics & Information Technology, 19: 107-115.

[17] LeCun, Y., et al., 2015, “Deep Learning”, in Nature, 521: 436-444.

[18] Meyden, R. V. D., Torre, L. V. D., 2012, Deontic Logic in Computer Science. Springer Berlin Heidelberg.

[19] Prakken, H.1997, Logical tools for modelling legal argument-a study of defeasible reasoning in law, Netherlands: Springer.

[20] Reiter, R.,1980, “A logic for default reasoning”, in Artificial Intelligence, 13:81-132.