Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 61–70

Procedia Computer Science

www.elsevier.com/locate/procedia

21st International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Optimal program for autonomous driving under Bentham- and Nash-type social welfare functions

Keita Kinjo[a,i], Takeshi Ebina[b]

[a]*Faculty of Economics, Okinawa International University, 2-6-1 Ginowan, Ginowan City, Okinawa, 901-2701,Japan*
[b]*Institute of Social Sciences, Shinshu University,3-1-1, Asahi, Matsumoto City, Nagano, 390-8621,Japan*

Abstract

The purpose of this paper is to formally define and solve ethical problems of how an artificial vehicle (AV) determines its driving behavior when there are some passengers in the AV and some pedestrians on a street. We construct a mathematical model introducing mainly two Bentham- and Nash-types social welfare functions, and derive optimal solutions. We show the optimal solutions are completely different depending on the functions and their parameters. Our contribution is that policymakers or managers of AVs can discuss the problem and determine an algorithm for autonomous driving by formalizing the situation and offering the optimal solutions.

Peer-review under responsibility of KES International

Keywords: Autonomous vehicle, Social welfare function, Ethics of artificial intelligence, Utilitarianism, Economics

1. Introduction

Artificial intelligence (AI) and robots are playing a more significant role in many situations today, from nursing to autonomous driving to military uses. Amidst this trend, decision-making and behaviors of AI and robots sometimes face social dilemmas from an ethical perspective, which is becoming a problem.[1,2] In particular, autonomous driving is known to have ethical issues.[3] It is a significant area of concern, and discussions are taking place on specific issues

[i] Corresponding author. Tel.: (+81)-98-892-1111; fax: (+81)-98-893-3273
E-mail address: keita.kinjo@okiu.ac.jp

1877-0509
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.024

and algorithms in this context.[4ii]

An example of a past study on autonomous driving is the decision-making problem investigated by Bonnefon et al. [5,6]. Based on an online survey, Bonnefon et al. analyzed an assessment of autonomous driving by people under three scenarios:

A) Consider a scenario in which a person crosses the road in front of an *Autonomous Vehicle* (*AV*). If the car continues to move forward, it will hit the person. On the other hand, if the car turns to avoid the person, it will crash into a wall, killing the passenger. How should the AV behave in this scenario? This is a problem in which the life of either the pedestrian or the passenger is prioritized over the other.

B) Next, consider a scenario in which many people are crossing the road, the automobile collides into a wall to avoid hitting the crowd, and the passenger dies. This is a problem in which the lives of either the crowd or the passenger is prioritized over the other.

C). Finally, consider a scenario in which many people are crossing the road in front of the car while it is driving autonomously. Turning the car to avoid the crowd results in running over another person. Such a scenario asks the question of whether the AV should drive straight or make a turn. This is a problem in which the lives of either the crowd or an individual are prioritized over the other upon comparing the lives of many against the life of an individual. Similar problems have been known for many years in the study of ethics, such as with the trolley problem.

These scenarios demonstrate the ethical problems that exist when automating cars using AI.[iii] Many studies have examined what choices people make under various ethical scenarios. However, few studies have discussed how the approaches to ethical problems should be defined formally, and how the optimal values vary in such formal settings. In particular, as the cases presented by Bonnefon et al. show, there is a strong need to solve problems under formal settings when ethical problems related to priorities in human damage are implemented in AI decision-making and in actual programs.

To formally define the ethical problems presented by Bonnefon et al., this study introduced social welfare functions used in economics as objective functions of AVs, and solved these functions. Finally, using the results, the impact of objective functions and changes in the parameters of control in AVs was discussed.

The problem raised by Bonnefon et al. is related to the problem of how social welfare should be defined. Several social welfare functions have been proposed in the field of economics. A social welfare function represents the sum of the welfare of individuals such as consumers, producers, and the government. Policymakers and social planners make decisions on policies to maximize this sum. The objective function for the policymakers is called a social welfare function. Some social welfare functions consider efficiency, while others consider equality. One way where economists might determine how best to balance competing objectives of efficiency and equality is to specify a social welfare function.[7,8] The details are discussed in the next section.

However, this study did not consider which function is preferred. This needs to be defined through surveys of the people, as studies by Bonnefon et al. have done. This is a matter of whether the passenger should choose upon purchasing the system, and whether it should be decided democratically upon designing the legal system.

This study found that optimal control varies by social welfare function in AVs. A major contribution of this study was the identification of an AV control for specific social welfare functions to enable the implementation of the program and a comparison of the results. Specifically, the study found that AV control under the Bentham welfare function (see the next section) becomes very extreme (boundary solutions), while the damages are halved by the control under the Nash welfare function (see the next section). The approach in this study proposed a roadmap to introducing the knowledge in welfare economics to autonomous driving and AI.

This paper is structured as follows. Section 2 introduces relevant studies, and model settings are defined in section 3. Optimal values are derived analytically per target function in section 4. Calculations are performed in section 5.

[ii] For example, Anderson [9,10] proposed a more specific algorithm to implement ethics in AI in general. There are systems such as "Jeremy" (based on utilitarianism) and W.D. (based on Ross's Moral Theory).

[iii] Similar problems arise in automation in marketing, in which utility for the people in business management is maximized. In marketing, segmentation is performed to maximize the utility for the target customers, and more effective actions are taken for specific customers. An example of this is a recommendation or visualization system. [11] For typical products, the problem of providing recommendations to specific customers for the purpose of efficiency does not occur. However, in cases where social assets are used (known as social marketing, for example, to promote medical exams), inequality arises when policies are specialized for certain active people. This is a problem that concerns the effective allocation of resources.

Section 6 discusses the results and future challenges.

2. Related studies

2.1. Introducing ethics to autonomous driving

A similar study was conducted by Gerdes et al.[12], who studied the approaches to integrating regulations and ethics that should be fundamentally observed in the control algorithms of AVs. Gerdes suggested that regulations and ethics can be introduced formally into autonomous driving as a cost function and constraint conditions for optimal control. In particular, the study used constraint conditions and soft constraint conditions to implement deontological ideas. Minimal rules that should be satisfied based on the Three Laws of Robotics postulated by science fiction writer Isaac Asimov were also defined.

Problem awareness in our study and that of Gerdes et al. is similar. However, our study is unique because it considers the priority in the utility of a person in scenarios where damages cannot be avoided, even after clearing the problem discussed above. In particular, our study is new because it focuses on the problem of utilitarianism in humans (which has been questioned by Bonnefon et al.), rather than on deontological issues.

2.2. Social welfare function

This section discusses major social welfare functions, beginning with the Bentham's utilitarianism.[13,14] This type of welfare function aims to maximize the sum of the welfare of the members that constitute a society. A major study by Bonnefon et al. approached the discussion based on utilitarianism. There is also a Nash study that maximizes the product of the welfare of the members.[15,16] Compared with Bentham's utilitarianism, the Nash product is an index that places more emphasis on equality. In Bentham's utilitarianism, since it is evaluated by the sum of the welfare of all individuals, welfare becomes sufficiently large if the welfare for one individual is large even when that of another individual is small. By contrast, the Nash solution measures social welfare by multiplying the welfare of the individuals, and the welfare thus becomes greater if the welfare values of the individuals are closer to each other. Therefore, it is considered an index that emphasizes equality. Rawlsian social welfare function pursues equality even more than Nash's social welfare function does. The solutions to ethical problems in autonomous driving can be aided by applying these functions.

3. Model

This section considers the problem of autonomous driving by formalizing a mathematical model. There are various types of problems in autonomous driving to consider, and it is difficult for researchers to comprehensively answer the ethical problems of autonomous driving. Hence, this study develops a model by modifying and restricting the situations that Bonnefon et al. consider in their paper[5] as a first step.[iv] Concretely, we focus on two of the three scenarios, (A) and (B), although Bonnefon et al. consider three scenarios, as explained in Footnote 2. Figure 1 shows an schematic of the stories.

Let the numbers of *passengers* and *pedestrians* be denoted as $n_a \in \mathbb{N}$ and $n_b \in \mathbb{N}$, respectively. The subscripts a and b represent passenger and pedestrian, respectively. $X \in [0,1]$ is a variable that controls the AV and represents the degree of the angle through which the steering wheel is turned; $X = 0$ holds if the AV swerves to avoid pedestrians and turns the steering wheel to the left, whereas $X = 1$ holds if the AV does not turn the steering wheel and continues moving forward. We also assume that the AV cannot move in the reverse direction and cannot turn the steering wheel to the right. In our setting, X affects the expected utilities of the passengers $U_a(X)$ and pedestrians $U_b(X)$, where

[iv] Bonnefon et al. consider the following three traffic situations: (a) the car can stay on course and kill several pedestrians or swerve and kill one passerby, (b) the car can stay on course and kill one pedestrian or swerve and kill its passenger, and (c) the car can stay on course and kill several pedestrians or swerve and kill its passenger.

$U_a: X \to \mathbb{R}$ and $U_b: X \to \mathbb{R}$ denote the utility functions of the passengers and pedestrians, respectively.[v] We also assume that the individual utilities of the passengers and pedestrians are the predicted and expected values and that these values are computable and comparable among them. For example, we can estimate these values using machine learning. Furthermore, we consider a situation in which the company of an AV or the policymaker constructs an algorithm for the AVs taking into account the utility functions of the passengers and pedestrians. The value of the car is negligible; hence, we can ignore the loss of the value when calculating the total value of an accident. We also assume that the passengers and pedestrians are immortal and do not die owing to the accident.

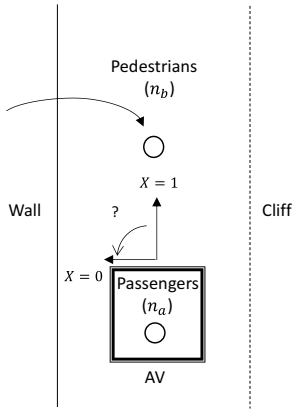

Fig. 1. Schematic of our situation.

Moreover, we assume that the value of the passengers is increasing with X, i.e., $U_a'(X) > 0$. This means that the utility of the passengers becomes smaller as the AV turns the steering wheel more to the left. In addition, we also assume that the value of the pedestrians is decreasing with X, i.e., $U_b'(X) < 0$. That is, if the AV turns the steering wheel more to the left, the utility level of the pedestrian becomes larger. Thus, the direction of the utility of the pedestrians toward the angle of the steering wheel is completely opposite to that of the passengers, and there exists a tradeoff between the pedestrians and the passengers with respect to the angle of the steering wheel. To simplify the mathematical analysis, the utility functions of the passengers and pedestrians are given as follows:

$$U_a(X) = \alpha X + c_a,\ (\alpha > 0, c_a > 0),$$

$$U_b(X) = -\beta X + c_b,\ (c_b > \beta > 0),$$

where the intercept of the utility function of the passengers, c_a, represents the utility level of the passengers when an accident occurs. The intercept of the utility function of the pedestrians, c_b, represents the utility level of the pedestrians when an accident does not occur. The slope of the utility function of the passengers, α, represents the additional disutility or additional cost by (decreasing X) steering, whereas that of the pedestrians, β, represents the additional disutility or additional cost by not (increasing X) steering.

Next, let us describe objective function W. First, the *Bentham-type social welfare function* that is often called utilitarian is defined as the sum of the utilities of all members, who are passengers and pedestrians, and is given as follows:

[v] Russel and Norvig[17] show that any rational agent can be described as possessing a utility function. They state that the utility function approach is useful, and provide speed and safety, as an example of elements of a utility function. Many researchers employ a utility-based agent and conduct research. Our model is developed in line with this stream.

$$\max_X W(X) = n_a U_a(X) + n_b U_b(X) \ subject \ to \ 1 \geq X \geq 0. \quad (1)$$

Second, the *Nash-type social welfare function* is defined as the product of the utilities of all members, who are passengers and pedestrians, and is given as follows:

$$\max_X \ W(X) = U_a(X)^{n_a} \times U_b(X)^{n_b} \ subject \ to \ 1 \geq X \geq 0. \quad (2)$$

Third, one considers two simple situations in which either the passengers or pedestrians maximize their utilities. The two maximization problems under constraints are given by

$$\max_X \ W(X) = U_a(X) \ subject \ to \ 1 \geq X \geq 0, \quad (3)$$

and

$$\max_X \ W(X) = U_b(X) \ subject \ to \ 1 \geq X \geq 0. \quad (4)$$

In the next section, we obtain the optimal solutions, compare them, and discuss how each parameter affects the optimal solutions depending on the four social welfare functions.

4. Optimal solutions of each social welfare function and comparison

This section presents the optimal solutions for the four different social welfare functions described in Section 3. First, regarding the Bentham-type social welfare function, solving the maximization problem in Eq. (1), we have Proposition 1.

Proposition 1. The optimal solution for the maximization problem in Eq. (1) under the Bentham-type social welfare function is

$X^{B*} = 1$ when $\alpha n_a - \beta n_b > 0$,

$X^{B*} = 0$ when $\alpha n_a - \beta n_b < 0$,

$X^{B*} \in (0,1)$ when $\alpha n_a - \beta n_b = 0$.

Proof of Proposition 1. Let us construct the following Lagrangian:

$$\mathcal{L}(X) = n_a(\alpha X + c_a) + n_b(-\beta X + c_b) + \lambda_1(1 - X) + \lambda_2(X - 0),$$

where λ_1 and λ_2 are Lagrange multipliers for the objective function. Differentiating $\mathcal{L}$ with respect to X, λ_1, and λ_2, there exist multipliers λ_1^* and λ_2^* such that

$$\frac{\partial \mathcal{L}}{\partial X}(X^{B*}, \lambda_1^{B*}, \lambda_2^{B*}) = (\alpha n_a - \beta n_b) + \lambda_1^{B*} - \lambda_2^{B*} = 0,$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_1}(X^{B*}, \lambda_1^{B*}, \lambda_2^{B*}) = 1 - X^{B*} \geq 0, \lambda_1^{B*} \geq 0, \frac{\partial \mathcal{L}}{\partial \lambda_1}\lambda_1^{B*} = 0,$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_2}(X^{B*}, \lambda_1^{B*}, \lambda_2^{B*}) = X^{B*} - 0 \geq 0, \lambda_2^{B*} \geq 0, \frac{\partial \mathcal{L}}{\partial \lambda_2}\lambda_2^{B*} = 0.$$

Therefore, solving these equations, we have (i) $X^{B*} = 1$, $\lambda_1^{B*} = (\alpha n_a - \beta n_b) > 0$, and $\lambda_2^{B*} = 0$; (ii) $X^{B*} = 0$, $\lambda_1^{B*} = 0$, and $\lambda_2^{B*} = -(\alpha n_a - \beta n_b) > 0$; and (iii) $X^{B*} \in (0,1)$, $\lambda_1^{B*} = 0$, $\lambda_2^{B*} = 0$, and $\alpha n_a - \beta n_b = 0$. Thus, we have the desired result. **(Q.E.D.)**

Proposition 1 states that the AV continues moving forward if the number of passengers or the slope of the utility function of the passengers is not very large; otherwise, the AV swerves and turns the steering wheel towards the wall. Note that if the ratio of the populations between the passengers and pedestrians, n_a/n_b, is equal to β/α, the AV determines the angle of the steering wheel to be between 0 and 1, $X^{B*} \in (0,1)$.

Next, we derive the optimal angle for the Nash-type social welfare function.

Proposition 2. The optimal solution for the maximization problem in Eq. (2) under the Nash-type social welfare function is

$X^{N*} = 1$ when $X_1^{N*} \geq 0$,

$X^{N*} = X_1^{N*}$ when $X_1^{N*} \in (0,1)$,

$X^{N*} = 0$ when $X_1^{N*} \leq 0$, where

$$X_1^{N*} \equiv \frac{\alpha c_b n_a - \beta c_a n_b}{\alpha\beta(n_a + n_b)}.$$

Proof of Proposition 2. Let us consider the shape of $W(X) = U_a(X)^{n_a} U_b(X)^{n_b} = (\alpha X + c_a)^{n_a}(-\beta X + c_b)^{n_b}$. Differentiating W with respect to X, the first-order condition is

$$\begin{aligned}\frac{\partial W}{\partial X} &= n_a\alpha(\alpha X + c_a)^{n_a-1}(-\beta X + c_b)^{n_b} - n_b\beta(\alpha X + c_a)^{n_a}(-\beta X + c_b)^{n_b-1} \\ &= (\alpha X + c_a)^{n_a-1}(-\beta X + c_b)^{n_b-1}[n_a\alpha(-\beta X + c_b) - n_b\beta(\alpha X + c_a)] = 0.\end{aligned}$$

Solving this first-order condition, the following three solutions satisfy the equality:

$$X = -\frac{c_a}{\alpha}, \frac{\alpha c_b n_a - \beta c_a n_b}{\alpha\beta(n_a + n_b)} (\equiv X_1^{N*}), \frac{c_b}{\beta}.$$

Because of the assumption of $c_b > \beta > 0$, we have

$$\frac{\alpha c_b n_a - \beta c_a n_b}{\alpha\beta(n_a + n_b)} - \frac{c_b}{\beta} = -\frac{\alpha c_b n_a + \beta c_a n_b}{\alpha\beta(n_a + n_b)} < 0,$$

and

$$-\frac{c_a}{\alpha} - \frac{\alpha c_b n_a - \beta c_a n_b}{\alpha\beta(n_a + n_b)} = -\frac{\alpha c_b n_a + \beta c_a n_a}{\alpha\beta(n_a + n_b)} < 0.$$

Thus, the relationships among the magnitudes of the three solutions are as follows:

$$-\frac{c_a}{\alpha} < X_1^* < \frac{c_b}{\beta}, \quad -\frac{c_a}{\alpha} < 0, and\ 1 < \frac{c_b}{\beta}.$$

Furthermore, the second derivative of W with respect to X at $X = X_1^{N*}$ is

$$\left.\frac{\partial W^2}{\partial X^2}\right|_{X=X_1^{N*}} = -\left[\frac{(\beta c_a+\alpha c_b)n_a}{\beta(n_a+n_b)}\right]^{n_a-1}\left[\frac{(\beta c_a+\alpha c_b)n_b}{\alpha(n_a+n_b)}\right]^{n_a-1} a\beta(n_a+n_b) < 0.$$

Therefore, we have shown that W has a local maximum at $X = X_1^{N*}$. If $X_1^{N*} \in [0,1]$, since W has a maximum value at $X = X_1^{N*}$, $X^{N*} = X_1^{N*}$ holds. On the contrary, if $X_1^{N*} > 1$, an extreme value does not exist for $X \in [0,1]$, and $W'(0) = c_a^{n_a}c_b^{n_b}(\alpha n_a/c_a - \beta n_b/c_b) > 0$ holds. Because the social welfare function is monotonically increasing in $X \in [0,1]$, $X^{N*} = 1$ holds. If $X_1^{N*} < 0$, an extreme value also does not exist for $X \in [0,1]$, and $W'(0) = c_a^{n_a}c_b^{n_b}(\alpha n_a/c_a - \beta n_b/c_b) < 0$ holds. Because the objective function is monotonically decreasing in $X \in [0,1]$, $X^{N*} = 0$ holds. **(Q.E.D.)**

Proposition 2 states that the optimal value of the steering angle depends on whether or not X_1^{N*} is between 0 and 1. The AV continues moving forward if $X_1^{N*} < 0$, whereas the AV swerves and turns towards the wall if $X_1^{N*} > 1$. On the contrary, if $X_1^{N*} \in [0,1]$, the AV turns the steering wheel and chooses an intermediate value, which is the significant difference between this Nash-type solution and the Bentham-type solution having the corner solutions, except for the special case where $\alpha n_a - \beta n_b = 0$ holds. Next, with respect to a property of X_1^{N*}, we have the following corollary.

Corollary 1. The AV controlled based on the Nash-type social welfare function is more likely to swerve and turn the steering wheel (decreasing X^{N*}) when (i) α, n_a, or c_b decreases and (ii) β, n_b, or c_a increases.

Proof of Corollary 1. Differentiating X^{N*} with respect to each parameter, we have

$$\frac{\partial X_1^{N*}}{\partial \alpha} = \frac{c_a n_b}{\alpha^2(n_a+a_b)} > 0, \quad \frac{\partial X_1^{N*}}{\partial \beta} = -\frac{c_b n_a}{\beta^2(n_a+n_b)} < 0, \quad \frac{\partial X_1^{N*}}{\partial n_a} = \frac{(\beta c_a+\alpha c_b)n_b}{\alpha\beta(n_a+n_b)^2} > 0,$$

$$\frac{\partial X_1^{N*}}{\partial n_b} = -\frac{(\beta c_a+\alpha c_b)n_a}{\alpha\beta(n_a+n_b)^2} < 0, \quad \frac{\partial X_1^{N*}}{\partial c_a} = -\frac{n_b}{\alpha(n_a+a_b)} < 0, \quad \frac{\partial X_1^{N*}}{\partial c_b} = \frac{n_a}{\beta(n_a+a_b)} > 0.$$

Therefore, X^{N*} is increasing in α, n_a, or c_b if $X_1^{N*} \in (0,1]$, whereas X^{N*} is decreasing in β, n_b, or c_a if $X_1^{N*} \in [0,1)$. Furthermore, if $X_1^{N*} \notin [0,1]$, X^{N*} has a corner solution, 0 or 1, and the parameters do not affect the optimal value X^{N*}. Therefore, we have the desired results. **(Q.E.D.)**

This corollary states the properties of how each parameter affects the optimal angle of the steering wheel. Concretely, as the number of passengers (pedestrians) decreases, the intercept of the utility function of the passengers (pedestrian) decreases, or the slope of the utility function of the pedestrians (passengers) decreases, then X_1^{N*} decreases, implying that the AV is likely to swerve and turn the steering wheel towards the wall. The result of Corollary 1 is very intuitive. The more the weight of the pedestrians (the number of pedestrians, the slope and intercept of the utility function of the passengers) outweighs that of the passengers, the more damage a group of passengers suffers from an accident.

Next, let us consider two situations in which the AV only considers the utility of the pedestrians in Proposition 3(i) and that of passengers in Proposition 3(ii).

Proposition 3. (i) The optimal solution when the AV only considers maximization of the utility of the pedestrians for the maximization problem in Eq. (3) is $X^{Pe*} = 0$.
(ii) The optimal solution when the AV only considers maximization of the utility of the passengers for the maximization problem in Eq. (4) is $X^{Pa*} = 1$.

Proof of Proposition 3. (i) Since $U_b'(X) < 0$, $n_b U_b'(X) < 0$ holds, implying that $X^{Pe*} = 0$.
(ii) Similar to (i), since $U_a'(X) > 0$, $n_a U_a'(X) > 0$ holds, implying that $X^{Pa*} = 1$. **(Q.E.D.)**

The results of Proposition 3 are very simple. When the AV only cares about one of the two groups, only one group is perfectly saved, whereas the other group completely suffers.

5. Numerical analysis

We conduct a numerical analysis to investigate how the parameters affect the optimal angle of the steering wheel. To simplify our analysis, we treat n_a and n_b as real numbers and not natural numbers. First, let us consider the effects of n_b on X^{B*} and X^{N*} in Propositions 1 and 2 and discuss how changing the ratio of the number of passengers to the number of pedestrians affects the two optimal values. Let us set the parameters as $\alpha = 1, \beta = 1, c_a = 2, c_b = 2$, and $n_a = 5$. This case is symmetric between the passengers and the pedestrians in a sense that $\alpha = \beta$ and $c_a = c_b$, except the ratio of the populations between the passengers and pedestrians, n_a/n_b, where n_a is fixed. The results are summarized in Figure 2.

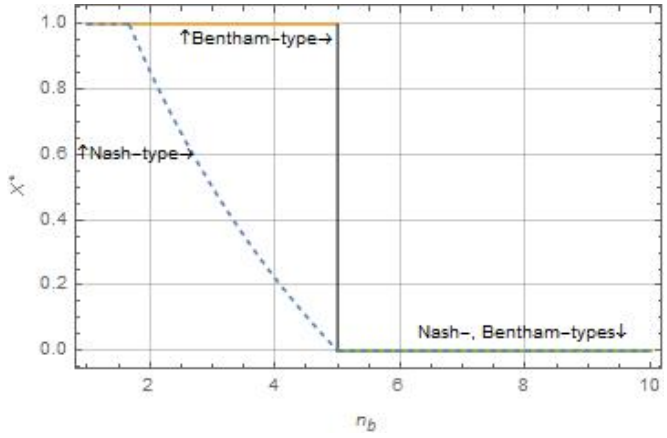

Fig. 2. Optimal angles for the Bentham-type (solid line) and Nash-type (dotted line) social welfare functions with respect to $n_b \in (0,10]$.

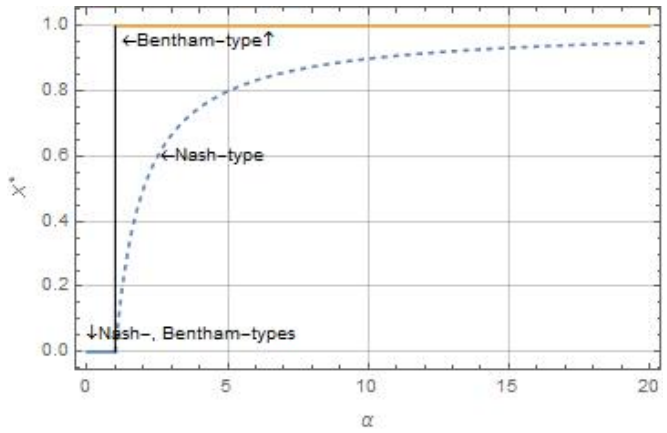

Fig. 3. Optimal angles for the Bentham-type (solid) and Nash-type (dotted) social welfare functions with respect to $\alpha \in (0,20]$.

Figure 2 shows that the AV drives into a wall if n_b is sufficiently large relative to $n_a = 5$ for both social welfare functions in the sense that $n_b \in (5,10]$ ($\Leftrightarrow X_1^{N*} \leq 0$) holds in this numerical example. However, the AV continues to move forward and strikes pedestrians for both social welfare functions if $n_b \in [1, 5/3)$ ($\Leftrightarrow X_1^{N*} \geq 1$) holds. In the two regions, $n_b \in [1, 5/3)$ and $n_b \in (5,10]$, the AV chooses the same behaviors. On the contrary, if $n_b \in (5/3, 5)$,

the optimal angles for the two social welfare functions are different. The optimal angle for the Nash-type social welfare function continuously changes as n_b changes, whereas for the Bentham-type social welfare function suddenly changes from 0 to 1.

Figure 3 shows another numerical example depicting the relationship between α and the optimal angles, X^{B*} and X^{N*}. Let us set the parameters as $\beta = 1, c_a = 2, c_b = 2, n_a = 5$, and $n_b = 5$. Note that X^{N*} converges to $X^{B*} = 1$ as α approaches infinity. Figure 3 exhibits similar properties in that the optimal angle for the Bentham-type social welfare function corresponds to that for the Nash-type social welfare function with respect to $\alpha \in [0,1)$ or $\alpha \to \infty$. On the contrary, unlike the previous numerical example in Figure 2, there exists a broad range of $\alpha \in [1, \infty)$ that does not match the two optimal angles.

From these two examples, one can first confirm that the opposite results do not emerge in the sense that neither $X^{B*} = 0$ and $X^{N*} = 1$ nor $X^{B*} = 1$ and $X^{N*} = 0$ hold simultaneously. This result states that the two social welfare functions have similar properties regarding the two parameters n_b and α. Let us note that the first result does not necessarily hold when c_b or c_b, which are not involved in the threshold of the Bentham-type optimal angle, $\alpha n_a - \beta n_b$, changes. Second, the optimal angle for the Nash-type social welfare function, X^{N*}, continuously changes, whereas that for the Bentham-type social welfare function, X^{B*}, jumps and switches from the one corner solution to the other. Thus, we can conclude that the Nash-type social welfare function distributes the loss of an accident more equally to both groups.

6. Discussion and concluding remarks

First, let us discuss the relationship between our results and the experimental results that Bonnefon et al.[5] obtained. Bonnefon et al. asked participants about their attitudes toward legally enforcing utilitarian sacrifices and pointed out that 76% of the participants thought that it would be more moral for AVs to sacrifice one passenger rather than kill 10 pedestrians. They also indicated that the same participants overwhelmingly expressed moral preferences for utilitarian AVs programmed to minimize the number of casualties. Our model suggests that when $n_a = 1$ and $n_b = 10$, since $X^{B*} = 0$ holds, we consider that $\alpha - 10\beta < 0 \Leftrightarrow \alpha < 10\beta$ also holds at the same time. In this situation, the results of this study are consistent with those of Bonnefon et al.. In addition, to compare our results with those of Bonnefon et al., let us set $n_a = 1$. Then, $X^* = 1$ holds when $\alpha > n_b\beta$, whereas $X^* = 0$ holds when $\alpha < n_b\beta$. In other words, if the number of pedestrians is larger than α/β, the AVs avoid the pedestrians. This result can capture a fundamental property behind the results of Bonnefon et al..

Next, let us summarize and discuss the setting and results of this study. This study considers the problem of how an AV determines its driving behavior when there are some passengers in the AV and some pedestrians on a street. The pedestrians suffer a monetary loss if the AV decides to continue moving forward, whereas the passengers suffer a monetary loss if the AV decides to swerve and crash into a wall. As the AV steers more to the left, the passenger suffer a greater monetary loss.

We have shown the optimal angles for the Bentham-type and Nash-type social welfare functions. We have also shown the qualitative properties for the two optimal values. For the Bentham-type social welfare function, the optimal value of the steering angle becomes positive or negative depending on whether $\alpha n_a - \beta n_b$ is positive or negative. This optimal value has two polar values, $X^* = 0$ or 1, except for a special case where $\alpha n_a - \beta n_b = 0$. Thus, the absolute values of the slope and the numbers of passengers and pedestrians are critical to determine how to drive an AV for the Bentham-type social welfare function. This is because the Bentham-type social welfare function is defined as the sum of the individual utilities and places a considerable amount of value on the efficiency of a society. For the Nash-type social welfare function, the optimal angle has an intermediate value in a broad range of the parameter spaces $(\alpha, \beta, c_a, c_b, n_a, n_b)$, which is completely different from the results for the Bentham-type social welfare function.

Our contribution is that policymakers or managers of AVs can discuss a problem and determine an algorithm for autonomous driving by mathematically formalizing the situation and offering the optimal solutions. In particular, our formalization enables researchers in the field of artificial intelligence or computer science to program a type of ethical problems, such as the driving of an AV. Moreover, our analysis suggests that the social welfare function approach used in economics has the possibility to be useful when applied to research on artificial intelligence.

Finally, let us mention the plans for future research. This study assumes that the utility functions are cardinal. This implies that one can compare the magnitudes of the utilities among people and carry out addition, multiplication, and so on. This setting is sometimes rather strong in economics because economists sometimes assume that the utility functions are ordinal. To apply the social welfare function approach, we need to examine this point further. Moreover, we exclude the problem of death of passengers and pedestrians when an accident occurs. However, when discussing a problem of autonomous driving, we cannot avoid a discussion on this type of problem. Hence, we also need to consider how the death of a person can be introduced into the social welfare function approach.

Acknowledgements

K. Kinjo acknowledges financial support from a Grant-in-Aid for Young Scientists (16K17203) from the Japan Society for the Promotion of Science. T. Ebina acknowledges a Grant-in-Aid for Young Scientists (15K17047) from the Japan Society for the Promotion of Science.

References

1. Deng B. Machine ethics: The robot's dilemma. *Nature* 2015;**523**(7558):24-6.
2. Kinjo K., Ebina T. Paradox of choice and consumer nonpurchase behavior. *AI & SOCIETY* 2015;**30**(2):291-7.
3. Greene JD. Our driverless dilemma. *Science* 2016;**352**(6293):1514-5.
4. Goodall NJ. Machine ethics and automated vehicles. In: Meyer G, Beiker S, editors. *Road vehicle automation*. Springer; 2014.
5. Bonnefon J-F, Shariff A, Rahwan I. Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? arXiv:1510.03346; 2015.
6. Bonnefon J-F, Shariff A, Rahwan I. The social dilemma of autonomous vehicles. *Science*, 2016;**352**(6293):1573-6.
7. Bergson A. A reformulation of certain aspects of welfare economics. *Q J Econ* 1938;**52**(2):310-34.
8. Samuelson PA. *Foundations of economic analysis*. Harvard University Press; 1947.
9. Anderson M, Anderson SL, Armen C. Towards machine ethics: Implementing two action-based ethical theories. *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics* 2005.
10. Anderson M, Anderson SL. Machine ethics: Creating an ethical intelligent agent. *AI Mag*. 2007;**28**(4):15.
11. Kinjo K., Ebina T. Case-based decision theory model matches ideal point model: application to marketing decision support system. *Journal of intelligent information systems*, forthcoming.
12. Gerdes JC, Thornton SM. Implementable ethics for autonomous vehicles. In: Maurer M, Gerdes C, Lenz B, Winner H, editors. *Autonomous driving*. Berlin Heidelberg: Springer; 2016. p. 87-102.
13. Arrow KJ, Sen A, Suzumura K, editors. *Handbook of social choice & welfare*, vol. 1. North Holland; 2002.
14. Arrow KJ, Sen A, Suzumura K, editors. *Handbook of social choice & welfare*, vol. 2. North Holland; 2010.
15. Mas-Colell A, Whinston MD, Green JR. *Microeconomic theory*. New York: Oxford University Press; 1995.
16. Kaneko M, Nakamura K. The Nash social welfare function. *Econometrica* 1979;**47**(2):423-35.
17. Russell S., Norvig P. *Artificial Intelligence (A modern approach)*.Prentice-Hall, 1995.