Ethics Inf Technol
DOI 10.1007/s10676-017-9419-3

ORIGINAL PAPER

A Rawlsian algorithm for autonomous vehicles

Derek Leben[1]

Abstract Autonomous vehicles must be programmed with procedures for dealing with trolley-style dilemmas where actions result in harm to either pedestrians or passengers. This paper outlines a Rawlsian algorithm as an alternative to the Utilitarian solution. The algorithm will gather the vehicle's estimation of probability of survival for each person in each action, then calculate which action a self-interested person would agree to if he or she were in an original bargaining position of fairness. I will employ Rawls' assumption that the Maximin procedure is what self-interested agents would use from an original position, and then show how the Maximin procedure can be operationalized to produce unique outputs over probabilities of survival.

Keywords Autonomous vehicles · Ethics · Rawls · Trolley problem

Introduction

Trolley dilemmas occur when a runaway object is headed towards some group of people, and it's possible to divert or block the train with an action that will result in the deaths of other people. Moral philosophers and psychologists are used to hearing this scenario criticized for being unrealistic and implausible. However, the newly emerging field of autonomous vehicles has recently provided a real-world trolley dilemma, and one that urgently needs to be solved before it happens, since the vehicles must be programmed with a decision-procedure prior to actually encountering the situation. The dilemma occurs when an autonomous vehicle is heading towards another vehicle or group of pedestrians, and does not have enough time to stop. In this situation, the vehicle can either swerve quickly, putting its own passengers at risk of harm, or slow down as much as possible while continuing its dangerous course.[1]

Most people seem to assume that some version of Utilitarianism provides the correct solution for this problem. In a recent survey of public responses, Bonnefon et al. (2016) found that laypeople overwhelmingly agree that swerving is the correct decision in the situation described above (less so when it's their own vehicle). Yet for those of us who believe that Utilitarianism is not the correct normative guide to decision-making, it is important to outline an alternative. One alternative would be a decision-procedure based on the Principle of Double-Effect: the vehicle always continues its deadly path towards pedestrians, on the grounds that this is merely foreseen (and regrettable) harm, while swerving to intentionally harm passengers would be intentional harm. This is not the alternative that I wish to present here. Instead, I will describe a decision-procedure for this problem using the principles and methods advocated by Rawls (1971).

Rawls' moral theory has been incredibly influential within moral philosophy since its original publication, but as far as I'm aware, it has never been operationalized for machine decision-procedures. In fact, the available texts on machine ethics (Wallach and Allen 2010; Lin

✉ Derek Leben
leben@pitt.edu

[1] University of Pittsburgh at Johnstown, 450 Schoolhouse Rd, Johnstown, PA 15904, USA

[1] More generally, the dilemma occurs whenever every action available to the vehicle will result in some amount of expected harm, whether this is from collisions with other vehicles, motorcycles, bicyclists, or pedestrians.

Published online: 04 March 2017

2011; Anderson et al. 2011) either don't mention Rawls or group him together with other theories like Kantian Ethics. In the past 15 years, there have been algorithms designed based on the principles of Utilitarianism (Anderson et al. 2004), Kantian Ethics (Powers 2006), Virtue Ethics (Wallach and Allen 2010), and Prima Facie Duties Approaches (Anderson and Anderson 2011). However, there have been no algorithms designed based on Rawls' version of Contractarianism, despite the fact that the theory contains what I take to be a very clear decision-procedure over numerical values. I also think that it happens to be the best available moral theory, but that would be too much to defend here. Instead, for the purposes of this paper, it will hopefully be enough to present the first description of a Rawlsian algorithm and demonstrate how it can be applied to autonomous vehicles.

The basic idea of this Rawlsian algorithm will be to gather the vehicle's estimation of probability of survival for each player in each action, then calculate which action each player would agree to if he or she were in an original bargaining position of fairness. I will employ Rawls' assumption that the Maximin procedure is what self-interested agents would use from an original position. This procedure will produce a unique decision in almost every situation, except the rare cases when there is a perfectly symmetrical trade-off of probability of survival for two or more players. Under these conditions, I suggest that self-interested players from the original position will randomize their decisions.

The next section of the paper will briefly describe some of the concepts and assumptions used in the algorithm (not to be confused with a detailed overview or defense of Rawls). The third section describes the details of the algorithm. The fourth section applies the algorithm to the problem of autonomous vehicles, and the fifth section presents some brief objections and replies.

A 'crash-course' in Contractarianism

Rawls' moral theory is usually categorized into the tradition of Contractarianism. The term 'Contractarianism' can refer to either a meta-ethical or a normative claim, and both will be adopted here. The meta-ethical claim is about the origin and purpose of moral rules; it states that moral principles are developed as solutions to the problem of social cooperation amongst self-interested organisms. The classic source for this view is Hobbes (1651), and has been endorsed in the twentieth century by Gauthier (1986), Binmore (2005), and Skyrms (2003). This view is described by Rawls in the opening pages of *A Theory of Justice*:

> Then, although a society is a cooperative venture for mutual advantage, it is typically marked by a conflict as well as by an identity of interests. There is an identity of interests since social cooperation makes possible a better life for all than any would have if each were to live solely by his own efforts. There is a conflict of interests since persons are not indifferent as to how the greater benefits produced by their collaboration are distributed, for in order to pursue their ends they each prefer a larger to a lesser share. A set of principles is required for choosing among the various social arrangements which determine this division of advantages and for underwriting an agreement on the proper distributive shares. These principles are the principles of social justice: they provide a way of assigning rights and duties in the basic institutions of society and they define the appropriate distribution of the benefits and burdens of social cooperation (Rawls 1971, p. 4).

The problem that Rawls describes here is prevalant in almost every part of human society. More food per person can be produced by agriculture than by hunting and gathering, but agriculture requires cooperation and storage which can be taken advantage of by thieves. Loaning money can be beneficial to both parties, but the lender always risks someone running off with her money. Two villages may gain strength by forming an alliance, but by letting their guards down, each runs the risk of its ally taking over. With the cooperators constantly putting themselves at such risk, one might expect that cooperation is a rare occurrence in animal societies. However, cooperation (both within and between species) is extremely common in both animal behavior and human societies. Thus, explaining how selfish organisms solve the problem of cooperation is a puzzle. Meta-ethical Contractarians suggest that moral principles and judgments have developed in order to overcome this challenge.

The best way of formally modelling the problem of cooperation employs games like the Prisoner's Dilemma (PD) and the Stag Hunt (SH), which both set up higher payoffs for cooperation than for working independently, but involve some risk in cooperation as well.[2] I'll call this category of games 'cooperation games.' In the PD, each player prefers mutual cooperation to mutual defection.

[2] The SH game comes from a story told by Jean-Jacques Rousseau about two hunters who could decide to either cooperate and hunt stag for a larger mutual payoff, or defect and decide to hunt hare for a lesser but still acceptable dinner (Skyrms 2003). The problem is that catching a stag requires two hunters, and so cooperating still makes the cooperator vulnerable. However, in this case (as opposed to PD), the other player doesn't have as much incentive to cheat, since a rabbit dinner could just as well be obtained from both players defecting.

Table 1 Payoffs in PD, where C = cooperate and D = defect

	C	D
C	2,2	0,3
D	3,0	1,1

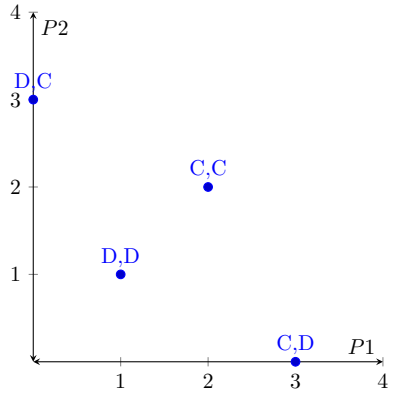

Fig. 1 Payoffs in PD as a graph, with P1's preferences on the *x*-axis, and P2's preferences on the *y*-axis

However, each player also most prefers to successfully take advantage of their partner, and least prefers to be taken advantage of. Using the conventions of game theory, this game is set up as a matrix with two players, where each player's preference is represented with an ordinal number, and higher ordinal numbers represent higher preferences (0 is one's lowest preference, 1 is preferable to 0, etc.) (Table 1):

One does not need to know the mathematical details of game theory to see that for both players, defecting is always the best option. If you are P1, you don't know what P2 is going to do: he might cooperate or he might defect. But you do know that if he cooperates, then your defecting will get you a payoff of 3 rather than 2. You also know that if he defects, then your defecting will get you a payoff of 1 rather than 0. The strategy of defecting in this game is called a strongly dominating strategy, because no matter what the other player does, this strategy always results in greater payoffs. This result is a surprise, since both players using their dominating strategies will produce the outcome of (1,1). However, there is another outcome preferred by both players: mutual cooperation (2,2). Thus, we find a paradox, where two self-interested organisms wind up producing an outcome that is not Pareto-optimal, meaning that there exists another outcome that is better for at least one player without making any others worse off.

The Contractarian meta-ethical proposal is that the function of morality is to push both players towards their Pareto-optimal outcomes. A system is Pareto-efficient (or Pareto-optimal) whenever no player's situation can be improved without making another player's situation somehow worse off. For example, you throwing away the rest of your dinner while I am sitting next to you starving to death would be Pareto-inefficient. Giving me your food would be a Pareto-improvement on the current state, since it would make me much better off, without making you any worse off (you were just going to throw the food away). If we are graphing the preferences of the two players on x- and y-axes, a Pareto-improvement on point *p* is just the set of all the points northeast of *p*. In Fig. 1, we see that mutual cooperation (2,2) is a Pareto-improvement on mutual defection (1,1) because it is northeast to the orthogonal lines dropped at (1,1).

The normative part of Contractarianism says that the best solution to the problem of social cooperation will ultimately be based in the (hypothetical or actual) agreement of players from some kind of idealized situation. Rawls' idealized situation is to cover up knowledge about oneself in a "veil of ignorance," which has the function of "[nullyifying] the effects of specific contingencies which put men at odds and tempt them to exploit social and natural circumstances to their own advantage" (Rawls 1971, p. 118). In this 'original positon,' players have information about the outcomes of each action without knowing which player they happen to be. The result of this is that all players will agree on the same decisions from the original position, since they are all effectively the same player.

Because the veil of ignorance puts an agent into a state of ignorance about her own position in the society, it is only applicable to determining the distributions of goods desired by every person to a comparable degree. For example, the original position would not be helpful in determining a fair distribution of calculators and lava lamps, because some people don't care about calculators at all. Instead, the original position method is limited to determining the distribution of what Rawls calls 'primary goods,' which are necessary requirements for the pursuit of any human interests. These include things like: life, opportunity, essential resources, and perhaps other things like health and survival. This is actually rather intuitive; the actions that people normally judge to be morally relevant are those that affect the distribution of opportunity, health, and essential resources, rather than those that affect the distribution of lava lamps. It's because autonomous cars will be making decisions that influence the distribution of health and survival that makes their decision-procedure morally relevant.

What decision-procedure will a rational person pursue from the original position? Imagine two potential distribution schemes of some primary good among a group of six people (for simplicity, we're assigning numerical values to the primary goods where higher numbers indicates more units of the good). Call these D_1 and D_2, where the two possible distributions are:

D_1: (0, 6, 7, 10, 20, 40)
D_2: (5, 6, 7, 7, 8, 9)

In the first scheme, Player 1 will receive 0 units of the primary good, and in the second scheme, Player 1 will receive 5 units of the primary good. Of course, from the original position, you don't know which player you will wind up being. If you're using a sum rule, you might prefer D_1, since it has a higher total number of units (83), which is higher than the sum of D_2. Similarly, if you're using an expected value rule, you'd still prefer D_1, since your expected value is roughly (.167)(83) = (13.83), which is higher than the expected value in D_2. This latter approach is advocated by the game theorist John Harsanyi (Harsanyi 1975). However, Rawls insists that, if you know the probability of being a particular person, you still aren't completely within the veil of ignorance, because your decisions won't be genuinely treating everyone's interests as potentially your own. Instead, he favors a 'thick' veil of ignorance (as opposed to Harsanyi's 'thin' veil), where the veil even covers information about how many people are in the society and how many are assigned each payoff.

Rawls claims that, within this thick version of the original position, every self-interested player will follow the Maximin criterion. Maximin is a strategy of maximizing the minimum payoffs; it is often described as being 'pessimistic' or 'safe,' because it focuses on improving the worst-case scenario. The strategy is usually employed within decision theory, where a player can make the decision that maximizes her lowest payoff (given the moves of other players). Within the original position, we have an equal chance of being the worst-off player as anybody else, so Maximin dictates that we maximize the minimum payoff for the entire set of players, rather than just for a single player alone. This is often described as making the worst-off person as well-off as possible. Accordingly, the principles of justice that Rawls eventually derives from the original position are focused on improving the welfare of the worst off people in a society. These principles for the distribution of rights, opportunities, and wealth will not be discussed in this paper, but only how the original position and Maximin procedure applies to decision-making in autonomous vehicles.

There is one part of the Maximin procedure that, to my knowledge, has not been worked out sufficiently by Rawls or anybody else, and is perhaps the only original contribution that I have to make to the moral theory itself. As I understand it, the Maximin procedure begins with the worst-off person in the set of players, but there is no reason in the framework of the original position why it has to stop with that person, why it would *only* apply to the worst-off person. It seems clear that agents in the original position would also consider the *next-lowest* payoffs, since they have an equal chance of being the next player, and are interested in maximizing her minimum as well. For example, consider two outcomes with payoffs of (2, 3) and (2, 90). Both of these outcomes are equivalent in their absolute lowest payoff, yet the second is obviously a Pareto-improvement on the first. I suggest that, from the original positon, once we have determined that these outcomes are equivalent in their absolute lowest payoff, we would mask the worst-off payoffs and perform the Maximin procedure on the next-lowest set of outcomes. In the present example, (2, 90) would clearly be the winner. Not only does this follow from the normative guidelines of the original position, but it now also satisfies the meta-ethical requirement of Contractarianism: results of the iterated Maximin procedure will always be Pareto-optimal. In addition, a unique outcome will be reached by this procedure in almost every situation, so we now have all the theoretical machinery needed to develop a decision-procedure for self-driving vehicles.

To wrap up this discussion of Rawlsian Contractarianism, let's see how the theory would apply to the PD game. For each outcome, we can create a set of the lowest payoffs, and then pick the outcome with the highest number in that set. In PD, the outcomes are: (1,1), (2,2), (0,3), and (3,0), so the corresponding set of lowest payoffs is: {1, 2, 0, 0}. The highest of these minima is the one corresponding to the outcome at (2,2), so that would be the outcome chosen by all players from the original position.

The algorithm

A model for the Rawlsian algorithm must contain at least three kinds of entities:

$$\text{N} = \text{Players} = \{P1, P2, P3, \ldots, Pn\}$$
$$\text{A} = \text{Actions} = \{A1, A2, A3, \ldots, Am\}$$
$$\text{u} = \text{Utility Function} = u : (A \times N) \rightarrow \{\mathbb{R}\}$$

Players are interpreted as all the people whose probability of harm is changed by the actions of the machine. Actions are interpreted as a set of outcomes that the machine can take steps towards at the present time. Rejecting the distinction between actions and omissions, we will also assume that do nothing is always an action within the set A.

The utility function is used to assign payoffs for each action with respect to each player. Formally, it is a mapping from the Cartesian product of players and actions to the set of real numbers. In the algorithm presented here, this function will be based on the probability that each player will survive, so all the numbers will be between 0 and 1. I'll assume that it is possible for an autonomous vehicle to estimate the likelihood of survival for each person in each outcome. Given that the vehicle will have information about its

Table 2 Action profiles as a table

	A1	A2	A3	A4
P1	.25	.10	.25	.30
P2	.70	.01	.10	.25

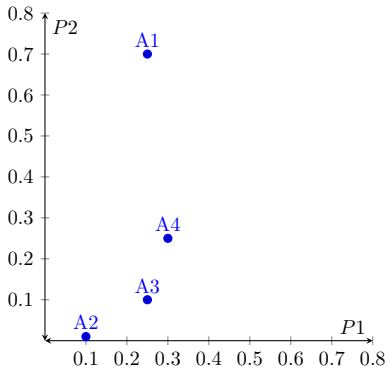

Fig. 2 Action profiles as a graph

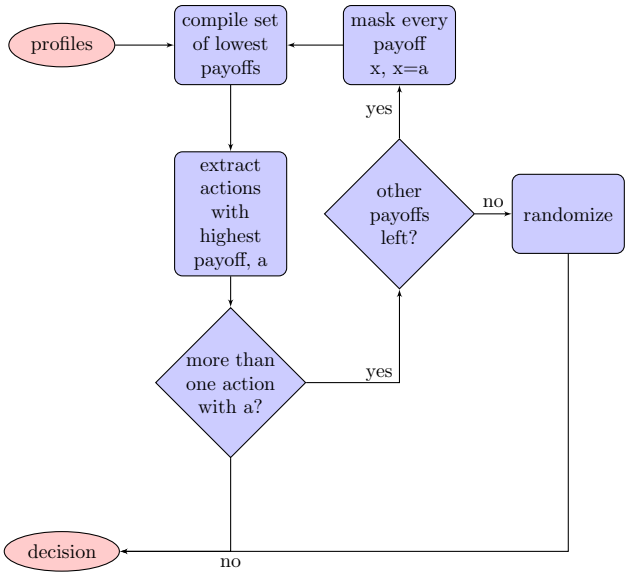

Fig. 3 The Rawlsian algorithm as a flowchart

own speed, the angle of impact, and sizes of the players, I think this is a plausible assumption to make. Indeed, one of the reasons why a Rawlsian algorithm may be successful in this domain is that goods in motor vehicle accidents can be considered along a single dimension of likelihood of survival. I'll also assume that injuries like broken ribs, whiplash, etc., can be represented as points along the dimension of likelihood of survival (the final section of this paper will consider a brief objection to this assumption).

The output of the utility function will be a set of action profiles, which assigns a real number to every pairing of players and actions. These action profiles can be represented as matrices (where each matrix contains the payoffs for each action), or as a table (Table 2) or a graph (Fig. 2):

Action profiles as a set of matrices:
A1: (.25, .70)
A2: (.10, .01)
A3: (.25, .10)
A4: (.30, .25)

Using the action profiles as input, a Maximin procedure can be run by compiling the lowest payoffs for each action into a set, then picking the highest payoff from that set. Call that payoff a. If a is unique in the set of lowest payoffs, then the procedure can halt and produce that action as its decision. For instance, in PD, the procedure would stop at this point, and produce mutual cooperation (2,2) as the result. However, if there is some other payoff in the set of lowest payoffs that ties for a, then it picks out more than one action. So the next step is to maximize the next-lowest payoff of those actions, assuming there are other payoffs besides a. This can be done by 'masking' (or deleting) the payoffs that are equal to a, $\{x|x = a\}$, and then running Maximin all over again on the remaining actions. We can continue doing this until there is a unique outcome, or we have run out of payoffs. At this point, self-interested players will be indifferent about the actions, and the procedure will randomize. This algorithm is represented in Fig. 3 as a flowchart:

I have designed this flowchart with typicality in mind. The three different columns (not including the input and output) are ordered from the most typical to the least typical paths. By far the most typical paths will trace a straight line down the first column and reach a single decision. Less typically, there will be 'ties' for the maximin value (a), and the path will proceed up through the second column and back down through the first again. In even more rare cases, this will continue in a circular path until a unique value is reached. Finally, in the rarest of circumstances (exactly symmetrical trade-offs in payoffs), the path will proceed to the third column and randomize.

Here are some examples. As described at the end of section , the Rawlsian solution to the PD game traces a straight path down the first column and reaches a unique output at (2,2). The data from Table 2 and Fig. 2 is more complicated; it proceeds up through the second column. For the action profiles A1-A4, the lowest payoffs are: {.25, .01, .10, .25}. The highest of these payoffs is .25, so well call that payoff a, and select every action with that payoff. In this case, both A1 and A4 have the same value a, so we select both actions, and proceed to the second column. Because there are still other payoffs in their action profiles, the

algorithm now masks *a* from both their profiles, and runs the Maximin procedure again. The new set of lowest payoffs (which is also the only set of payoffs left) becomes: {.70, .30}. The highest of these payoffs is .70; therefore, A1 is the decision output. If A1 and A4 had symmetrical payoffs in their outcomes, (.25, .70) and (.70, .25), then all the values would be masked and there would no longer be any payoffs left, so the procedure would move to the third column and randomize.

Applying the algorithm

We can now return to the original trolley-problem scenario for autonomous vehicles to see how the Rawlsian differs from the Utilitarian. I assume that a standard Utilitarian algorithm sums the payoffs in each action profile, and then selects the action with the highest total.[3] In many cases, these two algorithms agree. Consider actions A1 and A4 from the previous section, where the probabilities of survival are (.25, .70) and (.30, .25), respectively. Because these outcomes tied for the first round of Maximin, we masked the lowest payoff and found that A1 was the winner. The Utilitarian would agree, since the sum for A1 is .95, and the sum for A4 is .55. However, imagine now that the probability of survival for P1 in A1 is lowered even slightly to .24, so that the outcomes become: (.24, .70) and (.30, .25). Now, the Rawlsian algorithm proclaims A4 to be the winner, since it has the highest minimum payoff. However, A1 still has the highest sum by far, so Utilitarianism sticks by A1. The key conceptual difference is that the Rawlsian is unwilling to accept a lower opportunity for the worst-off player, even if it results in greater opportunities for everyone else.

In our model, the trolley problem can be constructed by letting A1 be the near certain death of the pedestrians and the near-certain survival of the vehicles passenger (continuing the present course), while allowing A2 to be its symmetrical opposite (swerving into an obstacle):

Classic Trolley:
A1 [continue]: (.99, .01, .01, .01, .01)
A2 [swerve]: (.01, .99, .99, .99, .99)

The Utilitarian algorithm will obviously favor A2, which has the highest sum of survival probabilities. On the other hand, the Rawlsian algorithm will see these as equivalent: after finding the same initial Maximin value ($a = .01$), the procedure will mask that value for all payoffs, and the next remaining value will also be identical (a = .99). After this, there are no longer any remaining values, so the vehicle will randomize. This makes sense conceptually: given that agents are choosing from behind the veil of ignorance, if I genuinely have no knowledge about whether I will be the passenger or the pedestrian, then I am genuinely indifferent about which action is better for me.

You might object that the Contractarian algorithm should not mask *all* the values equal to the minimum, but just one of them. If we did that, then Maximin would produce the Utilitarian solution (since the next-lowest values are higher for swerving than continuing). However, if we are using Rawls' 'thick' version of the original position, we should have no information about the number of people who are assigned to that outcome (and the probabilities of being these people). I'm only developing an algorithm based on this thick version, but it's certainly possible to develop a thin version of the Contractarian algorithm (which would still sometimes produce different results than the Utilitarian algorithm).

Importantly, the vehicle's decision changes immediately once the probabilities change even slightly. In one alternative scenario, it is more likely for the pedestrians to survive the vehicle hitting them than the passengers surviving a swerving move:

Trolley with Higher Payoffs for Pedestrians:
A1 [continue]: (.99, .05, .05, .05, .05)
A2 [swerve]: (.01, .99, .99, .99, .99)

In this scenario, the Rawlsian algorithm will now choose to continue its dangerous path, since A1 has the best minimum payoff ($a = .05$). In another scenario, the vehicle's passenger has a greater chance of survival in swerving than any of the pedestrians to by being hit:

Trolley with Higher Payoffs for Passenger:
A1 [continue]: (.99, .01, .01, .01, .01)
A2 [swerve]: (.04, .99, .99, .99, .99)

In this situation, the Rawlsian algorithm will swerve, since A2 has the best minimum payoff ($a = .04$). In all three of these scenarios, the Utilitarian will choose A2 (swerve), and the Double-Effect advocate will choose A1 (continue). However, the Rawlsian algorithm will change its decision depending on the minimum payoffs.

Objections and replies

In this section, I'll briefly consider three objections and replies. Before doing so, I'll say again that the purpose

[3] I am here ignoring the differences between Utilitarian procedures that sum the total and those that take an average (or weighted average). There are many sophisticated versions of the Utilitarian calculation, but I will only consider the most basic form here.

of this paper is not to defend Rawls' Contractarianism as a moral theory, but only to show how it can be developed as an algorithm and applied to the problem of autonomous vehicles. With that in mind, I won't be considering any of the standard objections to Rawls, but only problems that one might have with the way his theory is used here.

The first objection is that we should consider more than just survival probabilities when evaluating car crash outcomes. I should note, however, that I'm considering survival as part of (and a proxy for) physical health. Following Norman Daniels, I assume that health is a primary good that all people are interested in maximizing for themselves from the original position. Basically, I'm trying to set up a single scale of severe injury, with death being the most extreme point on this scale. I'm willing to revise this, and perhaps consider the most extreme kind of injury to be the most debilitating and painful injury that one could still survive, but it seems easier to use databases of fatalities and injuries that are more or less likely to lead to fatality.

You might be skeptical about setting up such a one-dimensional scale of health, with probability of survival being the proxy measurement. After all, survival might not have the same value when translated along the dimensions of age or social importance. A 50% probability of a 90-year-old person surviving might have a different value than a 50% probability of a 5-year-old surviving. A 50% probability of a person with an untreatable and terminal disease might have a different value than a 50% probability of a healthy person surviving. Even within the original position, there's a plausible argument that, not knowing which person you would be, any self-interested person from behind the veil of ignorance would favor the survival of someone with more 'quality-adjusted life years' (QUALYs) remaining in their lifetime over one with fewer (I'm aware that this is skating on thin ice, from a Rawlsian perspective). QALYs are a common tool for determining the allocation of scarce medical resources (Sassi 2006). If a computer were able to take into account the ages and health information of all the potential victims and run these into a function of remaining QALYs, then these values could be used as coefficients to weight the probabilities of survival. For example, if the probabilities of survival for Player A and B are: (.5, .8), but their remaining QALYs are: (30, 5), then their resultant values would be: ((.5)(30), (.8)(5)) = (15, 4). In this scenario, even though Player B has a higher probability of survival, Player A would have the higher health value when QALYs are factored into the calculation.

There is some reason to be hesitant about the use of QALYs in the Maximin calculation. There are already many moral objections to the use of QALYs to allocate scarce medical resources (Nord 1999), and I suspect that there would be public outrage to the idea that every person is attached a value based on their health that could be used to weigh human lives. As I hinted at, this is also skating on thin ice from a Rawlsian perspective, because it comes dangerously close to using social value as a factor in weighing lives. There are already rampant misunderstandings in the way that we should approach trolley problems: the MIT 'Moral Machine' game sets up trolley dilemmas where an autonomous car must choose between killing people based on information about the victim's weight, gender, employment, and criminal history. A Rawlsian (or any moral theorist, for that matter) would be aggressively opposed to this, for obvious reasons. Although I think the use of QALYs is still justifyable from the original position, it might be more politically risky than the use of bare survival probabilities.

The second objection to the algorithm presented here is that it will 'target' safer motorists in collision dilemmas, on the grounds that they have a higher probability of survival. For instance, if a vehicle must decide between colliding with two vehicles, and one has a higher safety rating than the other, the vehicle will pick the safer one (the occupant has a higher probability of survival). A recent article in *Slate* Magazine expressed fears about this:

> ...it seems unfair to penalize motorcyclists who wear helmets by programming cars to strike them over non-helmet wearers, particularly in cases where helmet use is a matter of law. Furthermore, it is good public policy to encourage helmet use; they reduce fatalities by 22-42 percent, according to a National Highway Traffic Safety Administration report. As a motorcyclist myself, I may decide not to wear a helmet if I know that crash-optimization algorithms are programmed to hit me when wearing my helmet. We certainly wouldn't want to create such perverse incentives.

First, it's misleading to say the algorithm is 'programmed to hit' anyone, just like it's misleading to call this 'targeting.' The algorithm would be programmed to avoid causing the lower minimum survival value. Using active terms like 'hit' or 'target' suggests that this is the goal or intention of the program. As for the concern that people are going to stop wearing helmets and buying safer cars, we should remember that dilemma situations are *extremely* rare; it's far more likely that a person will be involved in a normal collision than be involved in a dilemma-style collision. Any safety device also brings with it some small risk. Not wearing a helmet or buying a less safe car because you're worried about being targeted by an ethics algorithm would be like deciding not to wear a seat belt because they occasionally can lead to harm or death. Seat belts are much more likely to save you than kill you, although there is some small chance of the latter. Similarly, helmets and safe cars are much more likely to save you than result in you being targeted

by a crash optimization program, although there is some small chance of the latter.

The final objection is that the algorithm I've proposed will often produce counter-intuitive decisions. It should be noted, before I present the following scenario, that every algorithm based on a consistent moral principle will probably produce counter-intuitive decisions. With that being said, here is the most counter-intuitive one that the Rawlsian algorithm generates: imagine that a vehicle could either collide with a single pedestrian, causing almost certain death, or swerve into a crowd of pedestrians, causing many severe injuries. According to the Maximin principle, the car should swerve into the crowd. This is obviously surprising, and if we assume that the injuries are very severe (e.g., paralysis), you might even call this an insane decision. Technically, the Maximin principle prefers an infinite number of severe injuries to the death of a single person, and even non-Utilitarians might say this is a proof against the validity of the principle.

My first response is that this scenario is extremely unlikely; if a vehicle is equally unable to avoid hitting one pedestrian or another group of pedestrians, then the injuries both groups will sustain are probably going to be roughly equivalent. However, the abstract problem still remains. One shocking response is that, yes it seems counter-intuitive, but we're not relying on intuitions to generate or evaluate moral theories, so that's just a shame for our intuitions.[4] Another (less shocking!) reply would be to try and motivate this response a little bit to make it more intuitive. If I genuinely believe that I have an equal chance of being the person that dies, or one of the pedestrians that gets injured, then it does seem to me that I would always prefer to be one of the injured pedestrians (and thus I would prefer the action that produces this minimum outcome). If there are four passengers in the car, you could re-create the same objection: a crash that paralyzes all four passengers is not preferable to a crash that kills a single pedestrian. I find this intuitive as well, but I don't find it intuitive that it's better to paralyze a single person than to give four people broken legs. Perhaps if we remember that survival is being used as a proxy for health, with death being equivalent to the worst survivable injury, we can recover some of the intuitiveness of the Rawlsian prediction.

Conclusions and future work

This paper has presented a way of developing Rawls' Contractarian moral theory into an algorithm for crash optimization in autonomous vehicles. The proposed algorithm uses survival probabilities as data and produces a unique Pareto-optimal outcome based on the Maximin procedure and randomization. These results are importantly different from those produced by other theories like Utilitarainism. The chief advantage of a Rawlsian algorithm is its respect for persons as equals, and its unwillingness to sacrifice the interests of one person for the interests of others. Certainly, this can produce surprising results, but ones that any Rawlsian believes the foundations of morality must inevitably lead one towards.

I've been assuming that autonomous vehicles are operating more or less in isolation, but this is a simplification. Most major companies who are developing the technology (Google, Uber, Tesla) have their cars linked together and enable sharing of information. It's plausible that, in collision scenarios involving a set of autonomous cars that are networked together, there will be a different set of options available to the vehicle, and both vehicles will be able to make decisions simultaneously. For instance, if Vehicle A is going to collide with either a pedestrian or Vehicle B, then the two cars could simultaneously decide to adjust their positions in a way to optimize the resulting harm. Perhaps this would mean Vehicle B suddenly spinning to one side in order to avoid a head-on collision. While networked decision-making changes the actions and payoffs that are available to the vehicles, it doesn't change the basic algorithm that will be used to determine crash optimization. Both vehicles are still using either Maximin, the Utilitarain principle, or some other algorithm to determine which networked decision is best.

I have focused in this paper on applying the Rawlsian algorithm to autonomous vehicles, but there are many other domains where this algorithm would be relevant. Any machine whose actions influence the distribution of primary goods (health, survival, opportunity, essential resources) will face situations where it is impossible to avoid some kind of trade-off of these goods. This includes: military drones, personal assistants, home health care aids, and search-and-rescue robots. For example, if a search-and-rescue robot finds itself in a situation where it only has enough time or resources to save one group of people, it must make a moral decision. Assuming it's able to compute survival probabilities for each person in the group, it could run a number of procedures to make a decision (maximize the minimum, maximize the average, maximize the total), and the Rawlsian algorithm favors Maximin. Home care aids might face situations where two people are injured, perhaps two children have fallen off the playground, and it must attend to one of them first. Again, Maximin could be employed.

There are all sorts of practical problems that might arise in implementing this algorithm. It may turn out to be far

[4] This is not the response that Rawls would make, since he advocates a reflective equilibrium between our intuitions and our moral theories.

more difficult to estimate survival probabilities than I have assumed. Or it might take far too long to estimate survival probabilities and perform the Rawlsian calculation on them. The trolley-style problems discussed here require extremely fast decisions, and a solution that takes even a few seconds may be too long. However, this is an open question for computer science and engineering, and I am optimistic that these kinds of practical limitations are surmountable.

Just like Utilitarainism, Kantian Ethics, Virtue Ethics, and Prima Facie approaches can all be developed into algorithms for autonomous machine decision-making, I've tried to show that Rawlsian Contractarianism can also be translated into such an algorithm. The algorithm will turn out to be most valuable if, as I suspect, Rawls moral theory is correct. Nonetheless, even if Rawls moral theory turns out to be incorrect, it is still a valuable contribution to the field of machine ethics to show how different moral theories can be operationalized to produce different results.

References

Anderson, M., Anderson, S. L., & Armen, C. (2004). *Towards machine ethics*. AAAI-04 Workship on Agent Orientations: theory and practice.

Anderson, M., Anderson, S., & Leigh, S. (2011). *Machine ethics*. Cambridge: Cambridge University Press.

Anderson, S. L., & Anderson, M. (2011). A prima facie duty approach to machine ethics and its application to elder care. Human-Robot Interaction in Elder Care: Papers from the 2011 AAAI Workshop (WS-11-12).

Binmore, K. (2005). *Natural justice*. Oxford: Oxford University Press.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? *Science*, *352*, 1573–1576.

Gauthier, D. (1986). *Morals by agreement*. Oxford: Oxford University Press.

Harsanyi, J. (1975). Can the maximin princple serve as a basis for morality? A critique of john rawls' theory. *The American Political Science Review*, *69*, 594–606.

Hobbes, T. (1651). *Leviathan*. New York: Penguin Books.

Lin, P. (2011). *Robot ethics: The ethical and social implications of robotics*. Cambridge, MA: MIT Press.

Nord, E. (1999). *Cost-Value Analysis in Health Care*. Cambridge, MA: Cambridge University Press.

Powers, T. (2006). Prospects for a kantian machine. *IEEE Intelligent Systems*, *21*, 46–51.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Sassi, F. (2006). A prima facie duty approach to machine ethics and its application to elder care. *Health Policy and Planning*, *21*, 402–408.

Skyrms, B. (2003). *The stag hunt and the evolution of social structure*. Cambridge: Cambridge University Press.

Wallach, W., & Allen, C. (2010). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.