Blame in Autonomous Vehicle Collisions

Geoff Keeling[1]

Abstract. Drivers sometimes encounter *Catch-22 Collisions*, where they are faced with $n \geq 2$ decision-alternatives, and on each alternative, death or serious harm to at least one person is highly probable. In the near future, we can expect fully-autonomous driverless cars to be on our roads. It is plausible that driverless cars will, on occasion, face Catch-22 Collisions. How should manufacturers programme driverless cars respond to these collisions? And under what conditions are manufacturers blameworthy for harm caused in these collisions? In this paper, I articulate these two questions, and provide a partial answer to the second question.

1 INTRODUCTION

Drivers sometimes face ethical decisions. We can imagine a driver who must choose between exposing herself to a risk of death, or exposing another road-user to a risk of death. We can also imagine a driver who must choose between hitting five children at a crossing, or swerving and hitting one pedestrian on the pavement. In both cases, the driver must choose between $n \geq 2$ decision-alternatives, and on each alternative, at least one person is likely to suffer death or serious harm. I call these decisions *Catch-22 Collisions*.

We can expect driverless cars to populate our roads in the near future, perhaps as soon as 10 to 15 years. There is no good reason to assume that driverless cars will never encounter Catch-22 Collisions. We must therefore address the following ethical questions:

The Decision Problem: How should driverless car manufacturers programme driverless cars to respond to Catch-22 Collisions?

The Blame Problem: Under what conditions, if any, are driverless car manufacturers blameworthy for harm caused by driverless cars in Catch-22 Collisions?

In this paper, I articulate the *Decision Problem* and the *Blame Problem*. I then provide a partial answer to the *Blame Problem*. In Section 2, I explain the two problems and the difficulties we shall face in answering them. I then argue that we have good methodological reason to focus our initial efforts on the *Blame Problem*. In Section 3, I provide a partial answer to the *Blame Problem*. I argue that the conditions under which manufacturers are blameworthy in Catch-22 Collisions are different, in at least one respect, to the analogous blame-conditions for drivers in non-autonomous vehicles. Whilst drivers can be excused from blame in some cases because it is not reasonable to expect a driver to expose themselves to a significant risk of death to save another, manufacturers cannot be excused for this same reason. In Section 4, I consider the practical implications of this argument for driverless car regulators and manufacturers. In Section 5, I conclude.

2 TWO ETHICAL QUESTIONS

2.1 Assumptions

This paper is written conditional on some assumptions about driverless cars. Let us make these explicit.

I use the term 'driverless car' to mean either SAE Level 4 or Level 5 vehicles [1]. Level 4 vehicles are capable of full-automation in some locations, which means the car can drive in at least some areas without the need for human intervention or supervision. Level 5 vehicles are fully-autonomous, such that the car does not require human intervention or supervision in any location. In principle, Level 5 vehicles do not require human controls such as a steering wheel or brake-pedal.

Given the intentions of several manufacturers to produce driverless cars within the next 10 to 15 years,[2] we are justified in assuming

(A) Driverless cars will be on our roads soon.

Because the passengers in driverless cars will not be required, or in some cases be unable, to take control of the vehicle when Catch-22 Collisions arise, we are also justified in assuming

(B) Driverless cars must be programmed to respond to Catch-22 Collisions.

These assumptions have been defended elsewhere [2, 3, 4]. I shall take them for granted. Conditional on (A) and (B), we have good reason to ask how driverless cars should be programmed to respond to Catch-22 Collisions, and the conditions under which driverless car manufacturers are blameworthy for harm caused in these cases. I address each question in turn.

2.2 The Decision Problem

How should driverless cars be programmed to respond to collisions where harm to someone is inevitable? I believe there are, at least, three core difficulties that we shall face when providing an adequate answer to the *Decision Problem*. I consider each in turn. First,

(C) Proposed solutions to the *Decision Problem* are vulnerable to the same objections as the moral principles in which they are grounded.

I shall illustrate this point by example. Some philosophers have proposed utilitarian solutions to the *Decision Problem* [5]. There is also some empirical evidence purporting to show that

[1] Dept. of Philosophy, Univ. of Bristol, BS8 1TH, UK. Email: geoff.keeling.2016@my.bristol.ac.uk.

[2] See Google (Waymo Project); Ford (Smart Mobility Project); Volvo (Drive-Me Programme); Mercedes Benz (Model F 015); Nissan (ProPILOT 3.0 software); Rolls Royce (Model 103 EX); Tesla (Autopilot); Kia (Urban Autonomous Driving project) and BMW (Model iNext).

the public favours a utilitarian solution [6]. Let us suppose that a solution to the *Decision Problem*, *S*, is utilitarian just in case *S* mandates that manufacturers are morally required to programme driverless cars to minimise loss-of-life in token collisions; or to programme driverless cars to follow some rules that, overall, minimise loss-of-life in Catch-22 Collisions.

If a utilitarian solution is correct, then manufacturers are morally required to programme driverless cars to kill their passengers if doing so means minimising loss-of-life under certain conditions. We might object that utilitarian solutions place an unreasonable demand on manufacturers to give equal weight to the lives of their passengers and to the lives of other road-users. Manufacturers, after all, have a vested interest in their passengers' safety. It might also be objected that utilitarian solutions fail to respect the separateness of persons, insofar as they mandate aggregating human lives in a cost-benefit sum, which some argue is morally inappropriate.

Both these objections are special cases of complaints raised against utilitarianism simpliciter [7, 8]. However, these objections serve to illustrate the more general point: For any solution to the *Decision Problem*, the solution is vulnerable to at least the same objections as the moral principle(s) in which the solution is grounded. It therefore seems that a compelling answer to the *Decision Problem* requires engagement with current discussions in normative ethics, and plausibly further progress in these discussions.

It would be a mistake to think, however, that normative ethical considerations are the only considerations relevant to the *Decision Problem*. Consider,

(D) The solution to the *Decision Problem* must be action-guiding.

Though the *Decision Problem* is related to the question of which moral principle(s), if any, are correct, these problems are not identical [9, 10]. Driverless car decision-making is constrained by, at least, imperfect information and limited computational power. It is unclear, for example, that driverless cars will have the relevant technical capabilities to minimise loss-of-life in collisions, or to process complex moral distinctions like the difference between killing and letting-die. Given these limitations, we must not get caught up with idealised solutions to the *Decision Problem*, where we assume driverless cars to have unrestricted technical capabilities. Our solution must be action-guiding, insofar as it provides implementable advice to manufacturers on how to programme their cars to respond to Catch-22 Collisions.

Sceptics might respond that (D) presents no problem. We find the ideal solution and implement it as best we can. But this response fails to consider the following problem of underdetermination: Suppose 'minimise loss-of-life in token collisions' is the correct solution to the idealised *Decision Problem*. There are different implementable decision-rules that respect this ideal solution. Should manufacturers programme driverless cars to minimise expected loss-of-life conditional on the information available? Or should manufacturers programme the cars to minimise loss-of-life using a minimax or least-regret decision-rule? Clearly, idealised solutions do little to restrict the set of implementable action-guiding solutions to the *Decision Problem*.

Given (C) and (D), the *Decision Problem* presents a difficult conundrum. It would be misguided to focus exclusively on the normative ethical part of the problem. But it would also be misguided to focus our efforts on developing implementable solutions to the *Decision Problem* at the expense of serious normative ethical theorising. A defensible solution to the *Decision Problem* requires engagement on both fronts.

I believe the *Decision Problem* presents at least one further difficulty. Consider,

(E) The solution to the *Decision Problem* ought to be justifiable to each affected party.

There is a concern that condition (E) reflects my own normative ethical bias, as opposed to (C) and (D) which hold irrespective of our first-order normative ethical stance. However, it seems to me that if our proposed solution to the *Decision Problem* mandates killing someone under certain conditions, then we owe that person a justification for the decision to kill *them*. It is worth making explicit an important asymmetry in our ability to justify killing affected parties.

Suppose, in purchasing a driverless car, I agree to be killed under some conditions. If I am then killed under the correct conditions, then I have no straightforward moral complaint, because I consented to being killed under those conditions. But our solution to the *Decision Problem* is likely to mandate killing pedestrians and passengers under some conditions too; and it is much harder to provide a justification for killing these parties, because they will not have consented to being killed under the relevant conditions.

In short, the task of justifying our solution to the *Decision Problem* is not resolved by consent. This might be sufficient to cover the owners of driverless cars, but the scope of moral patients in Catch-22 Collisions is broader than the set of driverless car owners.

To summarise the discussion of the *Decision Problem*, I have presented three obstacles that stand in the way of our finding a solution. These obstacles are not insurmountable. But it should nevertheless be clear that finding a solution to the *Decision Problem* requires serious normative ethical theorising, in addition to considerations about what driverless cars are capable of, and the kind of justification required for killing people in Catch-22 Collisions.

2.3 The Blame Problem

Under what conditions, if any, are manufacturers blameworthy for harm caused by driverless cars in Catch-22 Collisions? *The Blame Problem* requires some clarification before we can explore its difficulties.

First, it might be assumed that the conditions under which manufacturers are blameworthy for harm caused in Catch-22 Collisions follow straightforwardly from the correct solution to the *Decision Problem*. For example, if the solution to the *Decision Problem* is 'minimise loss-of-life in token collisions', then manufacturers are blameworthy in all cases where the driverless car fails to minimise loss-of-life in token collisions. I find this response to the *Blame Problem* unconvincing. It relies on a commitment to the following principle:

Blame-Wrongdoing Equivalence: If some moral requirement, *M*, is the correct normative standard for some agent, *A*, in a scenario, *S*, then *A* is blameworthy just in case *A* fails to perform actions mandated by *M* in *S*.

We cannot take for granted the plausibility of this principle, for at least the reason that it commits us to the claim that there is no such thing as blameless wrongdoing. Thus, whilst we cannot reject outright the possibility that the solution to the *Blame*

Problem follows straightforwardly from the solution to the *Decision Problem*, there is good reason to avoid committing ourselves to this claim at present.

Second, to make clear exactly what the *Blame Problem* is asking, consider the same question asked of a human driver in a non-autonomous vehicle. Consider,

> *Necessarily Fatal Collision:* Jane is driving in icy conditions. There is a cyclist in front, who skids on the ice, falls, and lands in the middle of the road. Jane can (1) swerve off the road, saving the cyclist but killing herself; or (2) hit the cyclist, killing them, but saving herself.

We can ask the conditions under which Jane is blameworthy for harm caused in this scenario. This question has, at least, three components. First, which actions are morally wrong? Candidate actions include, (1), (2), and the actions Jane might have performed which caused *Necessarily Fatal Collision* in the first place (such as driving too fast given the weather conditions). Second, in performing any of these actions, did Jane fall short of her obligations to other road-users? Obligations here include following the legal rules of the road, paying attention, not performing reckless manoeuvres, and so on. Third, if Jane fell short of any obligation(s), can Jane be excused from blame for failing to meet the relevant obligation(s)?

To illustrate: Suppose that Jane was driving too fast for the weather conditions. Let us grant that this action is wrong, in a moral sense. We might then blame Jane for this action, in virtue of the fact that she fell short of her obligation not to perform reckless manoeuvres, and there is no reason to think that Jane could be excused from blame (such as Jane having a heart attack at the time of the action). I do not think that (1), sacrificing oneself to save a stranger, is wrong in a moral sense. So, this is not the kind of action that could incur blame. Last, even if (2) is wrong, and Jane fell short of an obligation not to put other road users in danger, we might excuse Jane from blame because she cannot reasonably be expected to sacrifice her own life to save a stranger.

The *Blame Problem* asks these same questions, where the manufacturer is taken to be the moral agent in a Catch-22 Collision involving a driverless car.[3] Consider,

> *Necessarily Fatal Auto-Collision:* A driverless car is driving in icy conditions with one passenger on board. There is a cyclist in front. The cyclist skids on the ice, falls, and lands in the middle of the road. The car can (1) swerve to avoid the cyclist, saving the cyclist but exposing the passenger to a risk of death; or (2) hit the cyclist, exposing them to a risk of death, but saving the passenger.

The solution to the *Blame Problem* should therefore provide answers to the following: First, are any of the actions in *Necessarily Fatal Auto-Collision* wrong? Second, did the manufacturer fall short of their obligations to other road-users? Third, if the manufacturer fell short of their obligations to other road-users, can the manufacturer be excused from blame for some reason?

I hope that the *Blame Problem* is now clear. I shall now turn to the difficulties it faces. I consider three naïve responses to the *Blame Problem*, and show that each of these faces at least one serious objection. First consider the

> *No Blame Thesis*: Manufacturers are never blameworthy when driverless cars cause harm in Catch-22 Collisions.

It might be argued that because someone will be harmed whatever the car is programmed to do, manufacturers cannot be blamed for harm caused in these collisions. But suppose that in *Necessarily Fatal Auto-Collision*, the reason that the driverless car ended up having to choose (1) or (2) is that its software had misread the weather conditions and the car was driving too fast. It seems the manufacturer could, at least, be blamed for this software failure and the deaths that occur in consequence. Intuitively, the manufacturer's omission to ensure that such software failures of this kind could not happen is wrong. In failing to take sufficient precaution when testing their software, the manufacturer might be blameworthy for falling sort of their obligation to, where possible, prevent harm to road-users. And it seems there is no obvious reason why the manufacturer can be excused from blame here. Next consider the:

> *Restricted Blame Thesis:* Manufacturers are blameworthy for harm caused in Catch-22 Collisions where this harm is the causal consequence of at least one software error. But manufacturers are not blameworthy for harm caused in Catch-22 Collisions where the harm caused is not the causal consequence of a software error.

The counterexample to the *No Blame Thesis* relied on a software error which the manufacturers could be blamed for. It might be argued that, provided no software error occurs, the manufacturer cannot be blamed for harm caused in necessarily fatal collisions. However, the *Restricted Blame Thesis* implies either: (i) It is impossible for manufacturers to make the wrong decision in Catch-22 Collisions; (ii) The manufacturers have no obligations to other road-users when driverless cars are faced with Catch-22 Collisions; or (iii) The manufacturer can always be excused from blame for wrongdoing in Catch-22 Collisions.

I believe that (i), (ii) and (iii) are false. Consider,

> *Crossing Calamity:* The driverless car's brakes fail. The car must either kill its passenger and 10 schoolchildren at a crossing, or swerve off the road, killing its passenger and no schoolchildren.

If the driverless car were programmed to kill the passenger and the schoolchildren, this would be the wrong decision. The deaths of the schoolchildren are preventable, and the passenger would die on either decision-alternative. Intuitively, the manufacturer is obliged to save lives where possible. In killing 11 rather than 1, the manufacturer falls short of this obligation. And there is no obvious reason to think that the manufacturer could be excused for this wrongdoing.

The final naïve response to the Blame Problem is the:

> *Equivalent Blame Thesis:* The conditions under which a driverless car manufacturer is blameworthy for causing harm in Catch-22 Collisions are the same as the conditions under which a human driver is blameworthy for causing harm in Catch-22 Collisions.

The Equivalent Blame Thesis faces at least two major objections. First,

(F) The obligations that manufacturers have towards road-users are plausibly different to the obligations that drivers have towards road-users.

There are, at least, two arguments for (F) First, driverless cars have different, if not superior, sensory apparatus and decision-making abilities compared to human drivers. For example, driverless cars can detect stimuli using radar and cameras 360-

[3] This assumption might be contested. But none of the arguments here are invalidated by substituting the manufacturer for a more specific set of people involved in programming the car to respond to Catch-22 Collisions in a particular way.

degrees around the vehicle, whereas human drivers have a 'blind spot'. Likewise, driverless cars can make decision-theoretic calculations under risk and uncertainty, whereas humans are unable to do this in collisions. Second, manufacturers make decisions about the outcomes of Catch-22 Collisions in advance of the collision occurring. Human drivers make the decision in the moment. Thus, it seems at least plausible that manufacturers can reasonably be expected to make a justifiable moral decision in Catch-22 Collisions, whereas it is unclear whether human drivers can reasonably be expected to make a justifiable moral decision in such collisions.

If either of these arguments hold, then we have reason to think that the obligations that driverless car manufacturers have towards road-users are stronger than the obligations that human drivers have towards other road-users. By extension, it seems at least plausible that the conditions under which manufacturers can fall short of their obligations to road-users are a superset of the conditions under which human drivers can fall short of their obligations to other road-users. Given this, it follows that the conditions under which manufacturers are blameworthy in Catch-22 Collisions are plausibly non-equivalent to the conditions under which human drivers are blameworthy. Hence, the *Equivalent Blame Thesis* is plausibly false. Next consider,

(G) The conditions under which manufacturers can be excused from blame in Catch-22 Collisions are plausibly different to the conditions under which human drivers can be excused from blame in Catch-22 Collisions.

I shall consider this point in detail in Section 3. The crux of the claim is that, in *Necessarily Fatal Collision*, we can excuse Jane for killing the cyclist on the grounds that she cannot reasonably be expected to sacrifice her life to save a stranger. This excuse is not straightforwardly relevant to the manufacturer of the driverless car in *Necessarily Fatal Auto-Collision*, because the manufacturer is not liable to die in the collision. If there exists at least one condition whereby human drivers can be excused from blame and manufacturers cannot, it follows that the *Equivalent Blame Thesis* is false.

To summarise the discussion of the *Blame Problem*, I have shown that three naïve answers to the *Blame Problem* are plausibly false. We therefore have good reason to consider carefully the conditions under which manufacturers are blameworthy for harm caused by driverless cars in collisions where death or serious harm is inevitable.

2.4 Methodological Priorities

I believe we have good methodological reason to focus our initial efforts on the *Blame Problem*, as opposed to the Decision Problem. This is for two reasons.

First, we must solve both problems in a short timeframe. Our intuitions about the *Blame Problem* are plausibly clearer than our intuitions about the *Decision Problem*. Whilst we can provide examples of situations where manufacturers are straightforwardly blameworthy, our intuitions about the moral principles that should guide driverless car decision-making in Catch-22 Collisions are less clear. This is, in part, because our intuitions about first-order normative ethical theories are conflicting, whereas our ability to identify situations where moral agents fall short of certain obligations (and the reasons why agents might be excused for doing so) are clearer. Given this, I think we have a reasonable chance of providing, at least, a partial answer to the *Blame Problem* in the relevant timeframe, whereas I am less hopeful for the *Decision Problem*.

Second, the *Blame Problem* might help us answer the *Decision Problem*. My claim is not that the *Blame Problem* is prior to the *Decision Problem* in a metaphysical sense, i.e. the conditions under which an agent is blameworthy for actions determine the moral standards by which an agent should act. But I think there is some coherence between what an agent ought to do and the conditions under which an agent is blameworthy. If we have good reason to think that manufacturers are blameworthy if they fail to save their passengers at all costs, then we also have reason to believe that utilitarian solutions to the *Decision Problem* are incorrect. It seems to me that getting clear about the blame-conditions for manufacturers could plausibly strike-off some candidate solutions to the *Decision Problem*, making it a more manageable ethical challenge.

For these two reasons, I believe we have good methodological reason to address the *Blame Problem* first.

3 ANSWERING THE BLAME PROBLEM

I shall not attempt to provide a complete answer to the *Blame Problem*. Instead, I focus on one of its three components: The conditions under which manufacturers can be excused from blame. I argue that the conditions under which driverless car manufacturers can be excused from blame in Catch-22 Collisions are different, in at least one respect, to the analogous conditions for drivers in ordinary vehicles. Consider again,

Necessarily Fatal Collision (NFC): Jane is driving in icy conditions. There is a cyclist in front, who skids on the ice, falls, and lands in the middle of the road. Jane can (1) swerve off the road, saving the cyclist but killing herself; or (2) hit the cyclist, killing them, but saving herself; and

Necessarily Fatal Auto-Collision (NFAC): A driverless car is driving in icy conditions with one passenger on board, call her Sue. There is a cyclist in front. The cyclist skids on the ice, falls, and lands in the middle of the road. The vehicle can (1) swerve off the road, saving the cyclist but killing Sue; or (2) hit the cyclist, killing them, but saving Sue.

NFC and NFAC are analogous in at least two respects: First, in each case, there are two decision-alternatives, and on each alternative, exactly one person will die. Second, in each case, one potential recipient of harm is inside the vehicle, and the other is outside the vehicle.

NFC and NFAC are dissimilar in the following respect: In NFC, the decision-maker is a driver, Jane, who must choose between saving her own life or that of a stranger. In NFAC, the decision-maker is the manufacturer, who must choose to save the passenger, Sue, or a stranger, the cyclist. Given this, there is a distinction in the logic of the conflict cases that underlie NFC and NFAC. We can distinguish:

Logic-NFC There are two agents, *A* and *B*, and *A* must save exactly one of *A* and *B*.

Logic-NFAC There are three agents, *A*, *B* and *C*, and *A* must save exactly one of *B* and *C*.

Most philosophers would agree that it is morally permissible for an agent to save themselves, when forced to choose between their own life and that of a stranger [11, 12]. I call this the 'right

to self-preservation'.[4] Given Jane's right to self-preservation, one of the blame-conditions for Jane in NFC is that she can be excused from blame if she chooses to save herself, even if this means falling short of her obligation not to expose other road-users to a risk of death.

The logic of NFAC does not permit the same argument. It is not the case that saving the cyclist would require the manufacturer to sacrifice themselves, so the manufacturer's right to self-preservation is not relevant to NFAC. Had the manufacturer programmed the driverless car to kill the cyclist even if the car had no passengers on board, then the manufacturer would be straightforwardly to blame for the cyclist's death. It seems, then, that one condition which can excuse drivers from blame for causing harm is not applicable to driverless car manufacturers.

It might be objected that, true, the manufacturer's right to self-preservation is not relevant in NFAC. But the passenger Sue's right to self-preservation is relevant, and the relation between Sue and the manufacturer is such that the manufacturer can excuse blame for killing the cyclist given Sue's right to self-preservation. It does not seem altogether implausible to think that there exist some cases where one agent can excuse blame for causing harm given another agent's rights. However, I do not think that the relation between Sue and the manufacturer is sufficient to establish that the manufacturer can appeal to Sue's right to self-preservation to excuse blame in NFAC-type cases. I shall consider two arguments against my position. The first concerns

Ownership: Sue's ownership of the driverless car is sufficient to establish that the manufacturer can appeal to Sue's right to self-preservation to excuse blame for killing the cyclist.

The plausibility of this argument depends on the inference from:

(H) A person owns a driverless car; to

(I) The manufacturer can excuse blame for causing harm given that person's right to self-preservation.

However, the inference from (H) to (I) faces a plausible counterexample. Consider,

The Old Switcheroo: The cyclist is the owner of the driverless car and she allows Sue to have a test-ride. The cyclist is cycling in front, she skids, falls onto the ice, and the car must either (1) swerve, killing Sue, but saving the cyclist, or (2) hit the cyclist, killing her, but saving Sue. The car is programmed to (1) kill Sue to save its owner.

I remain silent on whether the manufacturer is blameworthy for killing Sue in these circumstances. But it seems counter-intuitive to argue that the reason the manufacturer is not blameworthy is the cyclist's right to save herself in interpersonal trade-off cases. Such an argument merits a complaint from Sue: Plausibly, the fact that the car was driving on Sue's behalf provides stronger moral reason for the manufacturer to appeal to Sue's right to self-preservation, rather than the cyclist's.

It is, at least, unclear whether the manufacturer has stronger moral reason to appeal to the cyclist's right to self-preservation (given ownership) or Sue's right to self-preservation (given the car was driving on her behalf). But a robust defence of the ownership argument requires a decisive reason in favour of choosing the cyclist's right to self-preservation, rather than Sue's, in *Switcheroo* cases like the one above. In the absence of such a reason, the *ownership* argument fails.

The second argument purports to show that because the driverless car is *driving on Sue's behalf*, the manufacturer can appeal to Sue's right to self-preservation to excuse blame for killing the cyclist in NFAC. The inference here is from:

(J) A driverless car is driving on behalf of a person; to

(K) The manufacturer can appeal to that person's right to self-preservation to excuse blame for causing harm in collisions.

I think this argument also faces a plausible counterexample. Consider,

Two Passenger Railroad Case: Georgina and Bob are in a driverless taxi. Georgina is heading to destination *D*, and Bob is heading to destination *D'*. Because the destinations are close to one another, Georgina and Bob agree to share a taxi. The car is therefore driving on behalf of both Georgina and Bob. The car drives over a railway crossing, when suddenly the barriers shut and the car is stuck on the tracks. There are three separate tracks, and the car is presently parked over the middle track. On this middle track, two trains are heading from the left and the right, and both Georgina and Bob are certain to die if the car stays put. If the car reverses, then a train on the track behind will smash into the car from the left, almost certainly killing Georgina, but possibly saving Bob. If the car drives forward, then a train on the track in front will smash into the car from the right, almost certainly killing Bob, but possibly saving Georgina.

This case is fanciful, but I believe the logic of the case is sufficient to undermine the *driving on the passenger's behalf* argument. The logic of this case is such that the manufacturer, *A*, must save exactly one of *B* or *C*, or both will die. It does not seem to be the case that *B* or *C's* right to self-preservation is relevant to this case. If A saves B by appeal to *B's* right to self-preservation, then *C* has a moral complaint: A could equally well have appealed to *C's* right to self-preservation, and there was no good reason to appeal to *B's* right instead. The same holds in reverse. Thus, it seems that the fact (J) that a driverless car is driving on a person's behalf is not sufficient to establish that (K) the manufacturer of the car can appeal to that person's right to self-preservation to excuse blame in Catch-22 Collisions. The driving on someone's behalf argument therefore fails.

Given the failure of *ownership* and *driving on the passenger's behalf* to establish that a manufacturer can appeal to another's right to self-preservation to excuse blame in collisions, I believe we are justified in thinking that the blame-conditions for drivers and manufacturers are different in at least one respect. There exists a condition under which drivers can be excused from blame, namely the driver's right to self-preservation, where this right cannot be used to excuse manufacturers from blame.

[4] This is distinct from the right to self-defence, which requires, at least, a perceived threat from another. The right to self-preservation is the right that an agent has to save themselves in cases where they must save themselves or a stranger. The legal analogue is derived from the defence of necessity, wherein a person can be excused for causing harm where not committing that harm would expose them to potential death or serious harm. The legal analogue is not the defence of self-defence, where a person can be excused for causing harm to a perceived aggressor.

4 APPLICATIONS

I now consider the implications of the argument in Sections 2 and 3 for driverless car regulators and manufacturers. There are, I think, two points worth mentioning.

First, there is a discussion in the legal literature over how, if at all, manufacturer's liability law can be extended to the case of driverless cars [13, 14]. Though we have good reason to get to the bottom of the problem, we must also consider manufacturer's liability in Catch-22 Collisions where the harm caused by the driverless car does not arise from a software failure. Standard manufacturer's liability law is not equipped for such cases. In order to develop a comprehensive regulatory framework for driverless car collisions, we must think carefully about the obligations that can reasonably be ascribed to manufacturers programming driverless cars to respond to Catch-22 Collisions, and the conditions under which manufacturers can be excused from blame when driverless cars cause harm in these cases.

Second, the argument in Section 3 plausibly has implications for the *Decision Problem*. It might be assumed that manufacturers are justified in programming driverless cars to save their passengers in collision scenarios which require saving either the passengers or some other group of road-users. I have not shown that this position is incorrect. However, I have raised a serious challenge to one line of argument that might attempt to establish this position. If the passenger's right to self-preservation is not straightforwardly relevant to driverless car collisions, then manufacturers looking to 'save the passengers' owe either a substantive objection to the argument in Section 3, or a justification for the 'save the passengers' solution that does not rely on the passenger's right to self-preservation.

5 CONCLUSION

In this paper, I explained two ethical questions relevant to driverless car collisions where at least some harm is inevitable: The *Decision Problem*, and the *Blame Problem*. I then provided a partial answer to the *Blame Problem*, arguing that the conditions under which manufacturers are blameworthy for harm caused by driverless cars in Catch-22 Collisions are different to the conditions under which drivers are blameworthy for causing harm in Catch-22 Collisions. I argued that whilst drivers can excuse blame for causing harm given their right to self-preservation, there is good reason to think that manufacturers cannot appeal to the right to self-preservation to excuse driverless cars causing harm in the same circumstances. I then considered the applications of this finding for regulators and manufacturers investigating the ethics of blame and decision-making in driverless car collisions.

ACKNOWLEDGEMENTS

I am extremely grateful to Richard Pettigrew, Farbod Akhlaghi-Ghaffarokh, Nemo D'Qrill, Niall Paterson and Pavan Chaggar for their extensive support, comments and encouragement.

REFERENCES

[1] On-road Automated Vehicle Standards Committee. "SAE J3016: Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems." SAE International.

[2] Lin, Patrick. "Why ethics matters for autonomous cars." Autonomous Driving. Springer Berlin Heidelberg, 2016. 69-85.

[3] Goodall, Noah. "Ethical decision making during automated vehicle crashes." Transportation Research Record: Journal of the Transportation Research Board 2424 (2014): 58-65.

[4] Goodall, Noah J. "Machine ethics and automated vehicles." Road vehicle automation. Springer International Publishing, 2014. 93-102.

[5] Greene, Joshua D. "Our driverless dilemma." Science 352.6293 (2016): 1514-1515.

[6] Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. "Autonomous vehicles need experimental ethics: are we ready for utilitarian cars?." arXiv preprint arXiv:1510.03346 (2015).

[7] Hooker, Brad. "The demandingness objection." The problem of moral demandingness: New philosophical essays (2009): 148-162.

[8] Hart, Herbert LA. "Between utility and rights." Columbia Law Review 79.5 (1979): 828-846.

[9] Goodall, Noah J. "Away from Trolley Problems and Toward Risk Management." Applied Artificial Intelligence 30.8 (2016): 810-821.

[10] Nyholm, Sven, and Jilles Smids. "The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?." Ethical Theory and Moral Practice 19.5 (2016): 1275-1289.

[11] Taurek, John M. "Should the numbers count?." Philosophy & Public Affairs (1977): 293-316.

[12] Parfit, Derek. "Innumerate ethics." Philosophy & Public Affairs (1978): 285-301.

[13] Marchant, Gary E., and Rachel A. Lindor. "The coming collision between autonomous vehicles and the liability system." Santa Clara L. Rev. 52 (2012)

[14] Jeffrey K. Gurney, 'Sue my car, not me: Products liability and accidents involving autonomous vehicles', Journal of Law, Technology & Policy, (2013)