Ethic Theory Moral Prac (2018) 21:413–427
https://doi.org/10.1007/s10677-018-9887-5

Legal Necessity, Pareto Efficiency & Justified Killing in Autonomous Vehicle Collisions

Geoff Keeling[1]

Accepted: 4 April 2018 / Published online: 25 April 2018

Abstract Suppose a driverless car encounters a scenario where (i) harm to at least one person is unavoidable and (ii) a choice about how to distribute harms between different persons is required. How should the driverless car be programmed to behave in this situation? I call this the moral design problem. Santoni de Sio (Ethical Theory Moral Pract 20:411–429, 2017) defends a legal-philosophical approach to this problem, which aims to bring us to a consensus on the moral design problem despite our disagreements about which moral principles provide the correct account of justified harm. He then articulates an answer to the moral design problem based on the legal doctrine of necessity. In this paper, I argue that Santoni de Sio's answer to the moral design problem does not achieve the aim of the legal-philosophical approach. This is because his answer relies on moral principles which, at least, utilitarians have reason to reject. I then articulate an alternative reading of the doctrine of necessity, and construct a partial answer to the moral design problem based on this. I argue that utilitarians, contractualists and deontologists can agree on this partial answer, even if they disagree about which moral principles offer the correct account of justified harm.

Keywords Legal doctrine of necessity · Autonomous vehicle ethics · Robot ethics

1 Introduction

Driverless cars will be on our roads soon (Litman 2014). Some of these cars will require no human intervention or supervision in *any* circumstances. These are the cars with which I am concerned in this paper.[1]

[1] I use 'driverless cars' to mean Level 5 autonomous vehicles under the Society for Automotive Engineers autonomous vehicle classification system (n.d.).

Geoff Keeling
gk16226@bristol.ac.uk

[1] Department of Philosophy, Cotham House, University of Bristol, Bristol BS6 6JL, UK

Suppose a driverless car encounters a situation where (i) inflicting death or harm on at least one person is unavoidable; and (ii) a choice about how to allocate death or harm between different persons is required (Lin 2016; Gerdes and Thornton 2016). What does morality require of driverless car manufacturers in these cases? How, morally, should these cars be programmed to allocate death or serious harm between different persons? I call this the *moral design problem*.

Filippo Santoni de Sio (2017) defends a *legal-philosophical* approach to the moral design problem. According to this approach, we ought to take the criminal law as a starting point for our moral theorising. The rationale behind this approach is that answers to the moral design problem are grounded in moral principles. But people disagree about which principles provide the correct account of justified harm; and this disagreement is likely to prevent us reaching a consensus on the moral design problem before driverless cars are made available. Santoni de Sio uses the legal-philosophical approach to formulate an answer to the moral design problem based on the legal doctrine of necessity. In this paper, I raise an objection to Santoni de Sio's answer, and then use the legal-philosophical approach to formulate a new answer which overcomes the objection.

In §2, I distinguish two readings of the legal-philosophical approach. On the *wide reading*, the aim is to use the law to resolve our disagreements about which moral principles offer the correct account of justified harm. On the *narrow reading*, the aim is to use the law to bring us to a consensus on the moral design problem despite our disagreement about which moral principles offer the correct account of justified harm.

In §3, I explain Santoni de Sio's answer to the moral design problem, which is based on the legal doctrine of necessity. I then argue that, on either reading of the legal-philosophical approach, Santoni de Sio's answer falls short of its aim. On the narrow reading, Santoni de Sio fails to explain why postulating an alternative account of justified harm will resolve our disagreements about which moral principles offer the correct account of justified harm. On the wide reading, he fails to explain why advocates of competing accounts of justified harm should accept *his* answer to the moral design problem, given that it appeals to moral principles which are not unanimously agreed upon.

In §4 and §5, I offer a partial vindication of the narrow reading. I provide an alternative account of the doctrine of necessity which is based on a restricted Pareto principle. What distinguishes this reading from Santoni de Sio's is that it concerns the conditions under which necessity provides a justification for harm, as opposed to the normative principles which it is based on. I then use the restricted Pareto principle to formulate a partial answer to the moral design problem which provides a justification for inflicting harm in some, but not all, of the collisions with which the moral design problem is concerned. I argue that the restricted Pareto principle can be defended from the perspectives of three different moral theories: utilitarianism, contractualism and deontology. In doing so, I show that advocates of these theories have reason to accept the restricted Pareto principle as a partial answer to the moral design problem, despite their disagreements about which moral principles provide the correct account of justified harm.

2 The Legal-Philosophical Approach

The legal-philosophical approach aims to overcome a methodological challenge. I first state the challenge. I then distinguish two readings of how the legal-philosophical approach aims to address the challenge.

The challenge: answers to the moral design problem are grounded in moral principles. For example, Bonnefon et al. (2016) articulate a *utilitarian* answer, according to which driverless cars should be programmed to minimise loss-of-life in collisions; and Leben (2017) defends a *contractualist* answer, according to which driverless cars should be programmed to bring about the allocation of harm that the affected parties in the collision would rationally consent to behind a veil of ignorance. There is disagreement about which moral principles provide the correct account of justified harm. That is, as a matter of empirical fact, people disagree about whether utilitarianism, contractualism, or another set of principles, correctly describes which harms are morally permitted in cases where harm cannot be avoided. Santoni de Sio claims that this disagreement is likely to prevent us from reaching a consensus about the moral design problem before driverless cars are made available. As he puts it:

> […] both lay people and philosophers disagree about what is morally prohibited, permissible or obligatory in scenarios where fundamental interests and values are at stake, so that neither experimental ethics nor philosophical ethics seem at the moment able to offer [an uncontested solution to the moral design problem] (Santoni de Sio 2017: 412).

How does the legal-philosophical approach aim to overcome this methodological challenge? There are two interpretations. According to the *wide reading*, the law can bring us to a consensus about which moral principles provide the correct account of justified harm. If there is a consensus about these principles, then we can apply them to the moral design problem, and the methodological challenge is overcome. In support of this reading, Santoni de Sio writes:

> The main methodological idea behind this approach is John L. Austin's (1956) suggestion that legal reasoning may be a sharp instrument of clarification of complicated philosophical questions. According to Austin, the reflections of lawyers – with their standing attention to real-life cases […] may offer a fresh start to address difficult philosophical problems. Whereas I do believe that looking for fresh solutions to new or hard ethical problems is ultimately a philosophical enterprise […] I also think that philosophical reflection may sometimes benefit from considering legal principles and norms (2017: 413).

There is a second reading of the legal-philosophical approach. According to the *narrow reading,* the law can bring us to a consensus on the moral design problem *despite* our disagreements about which moral principles give the correct account of justified harm. Here, the aim is not for the law to solve our disagreements about which moral principles give the correct account of justified harm. The aim is to use the law to bring us to a consensus on the moral design problem in the absence of any general agreement about which moral principles offer the correct account of justified harm. In support of this reading, Santoni de Sio writes:

> […] legal norms are often the result of a combination of abstract moral principles and practical considerations deriving from the close observation and comparative analysis of real cases; moreover, legal norms are often an explicit attempt to cope with the fact of disagreement about general normative principles by finding a "reasonable compromise between principles and interests in contrast" (Hart 1961: 128) (*Ibid.*).

3 The Doctrine of Necessity

In this section, I first outline the doctrine of necessity. I then explain how Santoni de Sio uses this doctrine to formulate an answer to the moral design problem. Finally, I argue that on both the narrow and the wide readings of the legal-philosophical approach, Santoni de Sio's answer to the moral design problem fails to achieve the aim of the approach.

3.1 The Doctrine of Necessity

What is the doctrine of necessity? In broad terms, necessity is a legal defence. It stipulates sufficient conditions for a defendant to be absolved of criminal liability. Necessity arises in situations where the defendant faces

> [a] choice between values, protected interests, etc., where one of the defining features is that the prospective defendant is free to choose which course to take (Bohlander 2006: 150-1).

The necessity defence recognises that, sometimes, a person will (i) pursue a course of action which is ordinarily regarded as criminal, but (ii) their action should not be regarded as criminal because the decision they faced made it *necessary* for them to violate the ordinary wording of the law (Arnolds and Garland 1974: 290; see also *Reninger v Fagossa* 1551). For example, in *Mouse's Case* (1608), a river barge was threatened by a storm. The passengers were in danger of drowning unless the barge's cargo was thrown overboard. A passenger started throwing cargo overboard. One of the items thrown was a box which belonged to a fellow passenger called Mouse. The box contained £113, and Mouse sued the passenger for the loss. However, the court found that the passenger was not liable for the damages, because he had had acted out of necessity: whilst it would ordinarily be a criminal offence to throw Mouse's box overboard, the passenger was forced to choose between the competing interests of saving human lives or saving Mouse's box; and faced with this decision, the act of throwing Mouse's box overboard could not be regarded as unlawful.

I shall make necessity more precise by contrasting it with a related defence called duress of circumstances. The kind of choice required for a necessity defence – one between competing values or interests – often arises in concert with extreme circumstantial pressure. But circumstantial pressures are not relevant to the doctrine of necessity. This is because the necessity defence absolves liability by providing a *legal justification* as opposed to a *legal excuse*. Duress of circumstances holds that in some situations, circumstantial pressures are sufficiently extreme such that (i) the law considers the defendant's actions legally impermissible, but (ii) does not hold the defendant legally responsible for her actions. Hence, duress of circumstances provides a legal excuse. In contrast, necessity holds that sometimes a defendant will (i) perform an ordinarily criminal action, but (ii) their action is not criminal because they faced a decision which made it necessary to break the ordinary wording of the law. As Arnolds and Garland put it,

> [T]he courts limit the defence of duress to fear of serious bodily injury or death and make the defence personal to the person threatened. It makes no sense, however, to put those restrictions on the defence of necessity, since necessity is a justification and not an excuse (Arnolds and Garland 1974: 290).

Importantly, it is not required that the defendant faces a choice where *whatever* action she pursues violates the criminal law. Necessity can provide a defence in cases where the defendant breaks the law to prevent some greater evil occurring, even though the defendant would not have broken the law in letting the greater evil occur (*Re A* 2001). So, on a first pass, necessity is a *lesser of two evils* defence: it holds that a defendant is justified in breaking the law to prevent a greater evil occurring. As we shall see in our discussion of Santoni de Sio's analysis of the doctrine, this lesser of two evils characterisation admits too much. But it will serve as a good starting point for our discussion, as Santoni de Sio's analysis is pitched in opposition to the lesser of two evils – or as he calls it, the *simple utilitarian* – reading of the doctrine.

3.2 Santoni de Sio's Analysis

Santoni de Sio (2017) provides an insightful treatment of the doctrine of necessity. He uses this analysis to formulate an answer to the moral design problem. Before outlining the central themes of his discussion, I make two clarifications.

First, Santoni de Sio's discussion is extensive. At times, it goes beyond the scope of the moral design problem. I shall not discuss the parts of his argument which pertain to property damage, as the discussion here is limited to interpersonal moral dilemmas involving harm to persons. Neither shall I discuss the issues arising out of (i) driverless cars being used as public service vehicles; (ii) driverless car manufacturers' duty of care to passengers; (iii) motor vehicle users' duty of care to road-users; and (iv) broader questions about whether driverless car collision algorithms should be regulated by a public authority. These considerations are important in determining the scope of the doctrine of necessity as a legal justification for harm in driverless car collisions. They are also morally important, insofar as the justifiability of harm caused in driverless car collisions might depend on the normative positions of the relevant parties and their relations to one another. But my aim is to address the issue that our different moral commitments prevent us reaching a consensus on the moral design problem. I will focus on the parts of Santoni de Sio's analysis where advocates of three of our best moral theories can plausibly reach some agreement. It is unclear whether there is scope for similar agreement on the relevance of (i)-(iv) to the moral design problem. So, though the account of justified harm defended here sets aside some considerations which might be relevant the moral design problem, this is required to emphasise a point of agreement between some of our best moral theories.

Second, Santoni de Sio's analysis is structured as follows: he first discuses two landmark cases, *R v Dudley and Stephens* (1884) and *Re A* (2001). Both place restrictions on the doctrine of necessity which conflict with a simple utilitarian reading of the doctrine. Santoni de Sio then examines the normative considerations which motivate these restrictions. As my critique of Santoni de Sio focuses on just two of these considerations, the problem of incommensurability and the right to life, I shall discuss only these.

I start with the landmark cases. In *Dudley and Stephens*, four sailors were cast adrift off the coast of Africa. After many days without food, the cabin boy fell sick, and two of the sailors – Dudley and Stephens – decided to kill the cabin boy and eat him. The sailors were then rescued. When they returned to England, Dudley and Stephens were put on trial for murder. They attempted a necessity defence: if they had not killed the cabin boy, they would have starved to death. Lord Coleridge agreed that, had they not killed the cabin boy, they would have died. But he rejected this fact as sufficient grounds for a necessity defence. Santoni de Sio claims that Lord Coleridge's judgement was based on the principle that 'no innocent life should be taken under any circumstances' (Santoni de Sio 2017: 415).

Santoni de Sio then argues that 'things have changed' since Dudley and Stephens. He cites Brooke LJ's judgement in *Re A* as evidence for this. In *Re A*, the judges had to decide whether to permit a team of doctors to perform an operation that would separate conjoined twins. If the operation was not performed, both twins were guaranteed to die within a short time period. If the operation was performed, the weaker twin was guaranteed to die in the process. Brooke LJ permitted necessity as a legal justification for killing the weaker twin. He then set out three conditions for a successful necessity defence to killing in similar circumstances:

> a) the person killed was here the one (involuntarily) impeding the survival of the other; b) the person killed would have certainly died anyways in a short time; c) the killing has been committed with the official permission of a public authority (Santoni de Sio 2017: 416).

If Santoni de Sio's interpretation of these cases is correct, then both *Dudley and Stephens* and *Re A* impose restrictions on the doctrine of necessity which are at odds with the simple utilitarian reading. Lord Coleridge appeals to the principle that no innocent life should be taken under any circumstances, which the utilitarian does not recognise. In *Re A*, even setting aside Brooke LJ's requirement for the killing to be permitted by a public authority, the conditions for a successful necessity defence are stricter than a utilitarian reading of the doctrine maintains. I now consider two normative considerations which, on Santoni de Sio's analysis, underpin these additional restrictions on the doctrine of necessity: the right to life and the problem of incommensurability.

I start with incommensurability. Santoni de Sio (Santoni de Sio 2017: 419) distinguishes three interpretations of the problem of incommensurability. On a *conceptual* reading, it is impossible to compare the value of different lives because there exists no objective metric for measuring the value of different lives. On an *epistemic* reading, it might be true that there exists an objective metric for measuring the comparative values of different lives, but agents forced to make the kinds of decision that the moral design problem is concerned with are unlikely to have the resources to make a sound evaluation. On a *normative* reading, it does not matter whether the comparative value of lives can be measured: each individual has a right to life, and this must be respected irrespective of how valuable that life is. So, what is incommensurable *here* presumably is the value of one person's right to life and the value of another person's right to life.

Santoni de Sio's (Santoni de Sio 2017: 419) second argument against the simple utilitarian reading concerns the primacy of the right to life in criminal law. Following Christie (1999), he argues that an innocent bystander's *right to life* has stronger legal force than an individual's *request* to be rescued. As such, the doctrine of necessity does not permit programming a driverless car to kill an innocent bystander who would otherwise not be involved in the collision, to save the lives of those in the vehicle.

Santoni de Sio (Santoni de Sio 2017: 426) then concludes with the beginnings of an answer to the moral design problem[2]:

1. [there] might in principle be circumstances where a vehicle may be programmed to kill (*Re A*).

[2] I include three of Santoni de Sio's (Santoni de Sio 2017: 428) nine conclusions. The remaining conclusions reference the legal issues excluded from the discussion at the start of this section.

2. Given the strong restrictions to the intentional killing of innocents outside self-defence (*Dudley and Stephens*), the problem of the incommensurability of values and the right to life of persons, a program that allows for a vehicle to systematically hit persons who wouldn't be involved in the accident but for the vehicle decision seems unacceptable.
3. Based on the current legal constraints on killing under necessity (*Re A*), the intentional programming of a [driverless car] to target another [driverless car] might in principle be permitted under very specific and complex circumstances which seem unlikely to be realised in the near future.

3.3 The Dilemma

I now raise a dilemma against Santoni de Sio's answer to the moral design problem. In §2, I argued that the legal-philosophical approach admits two readings. I argue that, on either reading, Santoni de Sio's answer to the moral design problem falls short of the aim of the legal approach.

I start with the wide reading. The aim of the legal-philosophical approach, on this reading, is to use the law to resolve our disagreements about which moral principles provide the correct account of justified harm. But what Santoni de Sio has done is provide an alternative account of justified harm which is based on the doctrine of necessity. Santoni de Sio's account stands in competition with existing accounts, such as utilitarianism and contractualism. He has not explained *why* advocates of competing views will accept his alternative. Indeed, there is good reason to think that advocates of at least one competing view will *not* accept his necessity-based account. Consider utilitarianism. Santoni de Sio's necessity-based account of justified harm is unlikely to convince utilitarians for two reasons. First, in some cases, it delivers the wrong verdict: Santoni de Sio takes some harms as impermissible which utilitarianism deems permissible (e.g. the killing of an innocent person who is *not* involuntarily impeding the survival of another, in order to save a greater number of lives).[3] Second, the approach is justified by appeal to two considerations which the utilitarian does not take as morally salient: the problem of incommensurability and the right to life. So, without an explanation of why the utilitarian should accept his competing account of justified harm, it is unclear how Santoni de Sio's necessity-based account can resolve our moral disagreements. In turn, it is unclear how Santoni de Sio can bring us to a consensus on the moral design problem.

I now address the narrow reading. The aim of the legal-philosophical approach, on this reading, is to bring us to a consensus on the moral design problem *despite* our moral disagreements. The first problem is the same as above. Why should utilitarians accept Santoni de Sio's necessity-based answer to the moral design problem given that it appeals to principles which they do not recognise? I suspect Santoni de Sio would argue as follows: the law reaches its conclusions through a combination of moral and practical reasoning, which allows us to reach definite normative conclusions in the absence of any agreement about which moral principles give the correct account of justified harm (Santoni de Sio 2017: 413).

But the utilitarian can respond: the moral design problem is a *moral* problem. Answers to the moral design problem must be justified with moral reasons. Perhaps the legal-philosophical approach can bring us to a consensus on an *all things considered* design problem, which factors

[3] Not all forms of utilitarianism permit such killings. But my argument succeeds provided there exists at least one plausible account of utilitarianism which would permit this.

in both moral and practical reasons. But the aim of the legal-philosophical approach is to bring us to a consensus on the *moral* design problem despite our disagreements about which moral principles provide the correct account of justified harm. Santoni de Sio needs to provide a *moral* argument for his necessity-based solution to the moral design problem. Furthermore, advocates of competing accounts of justified harm must have reason to accept this argument despite their disagreements about which moral principles are correct. Without this argument, it is unclear how Santoni de Sio's necessity-based solution can bring us to a consensus on the moral design problem which advocates of different moral principles can accept *despite* their disagreements.

In §4 and §5, I attempt to salvage the narrow reading from this objection. The objection to the narrow reading arises because Santoni de Sio provides an analysis of necessity based on the normative principles which underpin it. This is problematic because disagreement about normative principles gives rise to the problem which the legal-philosophical approach aims to overcome. I shall provide an alternative account of necessity, based on the conditions under which it provides a justification for inflicting harm, as opposed to the principles on which the justification is based. I argue that the conditions under which the doctrine applies are captured by a restricted Pareto principle. I use this principle to formulate a partial answer to the moral design problem. The answer is *partial* insofar as it provides an account of justified harm in some, but not all, of the collisions with which we are concerned.[4] I argue that this principle can be defended from the perspectives of three different moral theories: utilitarianism, contractualism and deontology. If I am correct, then the legal-philosophical approach can bring us to a consensus on a decision-rule to use as a partial answer to the moral design problem in the absence of any agreement about which moral principles provide the correct account of justified harm. So, whilst the law might not be able to bring us to a *complete* consensus on the moral design problem despite our moral disagreements, I believe it can bring us to a partial consensus.

4 An Alternative Reading of Necessity

I now present an alternative reading of the doctrine of necessity. I then use this to formulate a *partial* solution to the moral design problem. I start with a statement of the principle which, I argue, captures the conditions under which necessity provides a justification for inflicting harm. According to

> *The Restricted Pareto Principle (RPP):* In situations where (i) harm to at least one person is unavoidable, and (ii) a choice about how to allocate harm between different persons is required, then *if* there exists a unique Pareto efficient allocation of harm across different persons, *then* other things being equal,[5] bringing about the Pareto efficient allocation of harm is justified.

[4] I believe that the doctrine of necessity can *at most* provide a partial answer to the moral design problem. This is because the doctrine of necessity provides a justification for harm in some, but not all, interpersonal moral dilemmas, and the set of interpersonal moral dilemmas which could arise in driverless car collisions is a superset of those for which the doctrine of necessity provides a justification for harm.

[5] I stated earlier that my intention was to focus on a part of the doctrine of necessity where three of our best moral theories can reach some agreement. I have included an 'other things being equal clause' to leave open the possibility that other considerations based on the normative positions of the relevant actors (e.g. responsibility) might defeat the prima facie justification provided by the Restricted Pareto Principle. I am grateful to an anonymous reviewer for pressing me on this point.

Three clarifications: First, an allocation of harm is Pareto efficient if, and only if, there exists no alternative allocation of harm in which all affected parties are at least as well-off, and some affected party is strictly better off. Second, there is a *unique* Pareto efficient allocation of harm when there is exactly one Pareto efficient allocation of harm among the alternatives. Third, RPP is a *sufficient* condition on justified harm in the cases where it applies. It does not follow from RPP that it is unjustifiable to bring about an outcome which is *not* uniquely Pareto efficient in scenarios where (i) and (ii) obtain.

I have explained what RPP amounts to. I now defend RPP as a plausible reading of the doctrine of necessity. My argument draws on the distinction that Brooke LJ drew between *Re A* (2001) and *Dudley and Stephens* (1884) in *Re A*.

Consider again the facts of *Re A*. Gracie and Rosie were conjoined twins. The doctors faced two options: (1) perform the operation and separate the twins; or (2) do not perform the operation. The medical evidence suggested that, on (1), the probability of Gracie surviving was 90% and the probability of Rosie surviving was 0%. On (2), both were almost certain to die. Brooke LJ permitted the defence of necessity as a legal justification for the doctors' performing the operation and killing Rosie in the process. As Lord Coleridge denied necessity as a defence to killing the cabin boy in Dudley and Stephens, Brooke LJ needed to show that the facts of *Re A* were sufficiently different, such that the precedent in *Dudley and Stephens* did not preclude necessity as a justification for killing Rosie in *Re A*.

To understand Brooke LJ's distinction between the two cases, let us examine *why* Lord Coleridge disallowed necessity as a justification for killing in *Dudley and Stephens*. Consider,

> By what measure is the comparative value of live to be measured? Is it to be strength, or intellect, or what? It is plain that the principle leaves him who is to be profited by it to determine the necessity which will justify him in deliberately taking another's life to save his own. In this case the weakest, the youngest, the most unresisting, was chosen. Was it more necessary to kill him than one of the grown men? The answer must be "No" (1884: 287-8).

Lord Coleridge makes three points here. First, it is unclear how, if at all, we can measure the relative value of the lives of different persons. Second, if necessity could be used as a justification for killing in the circumstance of *Dudley and Stephens*, then it would be open to abuse. Third, that there was no good reason to kill the cabin boy instead of one of the other sailors.[6]

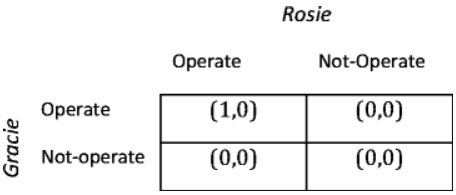
Brooke LJ distinguished *Re A* from *Dudley and Stephens* by considering two of these remarks. First, if Lord Coleridge had permitted necessity as a defence for killing the cabin boy, then the law would need to have weighed-up the comparative value of the sailors' lives against that of the cabin boy. However, in *Re A*, Rosie was guaranteed to die irrespective of whether the operation was performed. So, permitting necessity as a defence to killing in *Re A* would not require the law to weigh-up the comparative value of Gracie and Rosie's lives. Second, in *Re A*, the decision to kill Rosie instead of Gracie was not arbitrary: there was a good reason to kill Rosie, because she was impeding Gracie's survival, and Gracie was not impeding her survival. But in *Dudley and Stephens*, all the sailors were impeding each other's survival in the relevant sense: any one

[6] It should be noted that there was a reasonable expectation that the cabin boy would die sooner than the others because he had fallen sick. But it is clear from Lord Coleridge's judgement that this was insufficient reason to choose to kill the cabin boy over one of the other sailors.

could have been killed to save the other two. Brooke LJ clarified his distinction by providing an analogy. Consider,

> [T]he same considerations would apply if a pilotless aircraft, out of control and running out of fuel, was heading for a densely populated town. Those in the aircraft are in any event "destined to die". The would be no question of human choice in selecting the candidates for death […] if their inevitable deaths were accelerated by the plane being brought down on waste ground (*Re A* 2001: 85).

There is a game-theoretic similarity between the facts of *Re A* and the aeroplane example, and this similarity is not shared by *Dudley and Stephens*. Consider the situation of *Re A*, depicted in a payoff matrix:



		Rosie	
		Operate	Not-Operate
Gracie	Operate	(1,0)	(0,0)
	Not-operate	(0,0)	(0,0)

The outcome which obtains if Gracie and Rosie both choose Operate is Pareto efficient, because there exists no other outcome which makes it the case that both twins are at least as well-off and at least one twin is strictly better off. All the remaining outcomes are Pareto inefficient. So, *Re A* had a *unique* Pareto efficient outcome, where both Gracie and Rosie 'agree' to perform the operation. In *Dudley and Stephens*, granting that at least one person had to be killed if any of the parties were to survive,[7] there is no unique Pareto efficient outcome. If Dudley and Stephens kill the cabin boy, this outcome is Pareto efficient, because there is no other outcome such that all three are at least as well off and at least one is strictly better off. The same holds if Dudley and the cabin boy kill Stephens, or if Stephens and the cabin boy kill Dudley.

Thus, at least one salient feature of Brooke LJ's distinction between *Re A* and *Dudley and Stephens*, is that *Re A* had a unique Pareto efficient outcome and *Dudley and Stephens* did not. I therefore take RPP as a plausible account of the conditions under which necessity provides a justification for harm.[8] We can use this to formulate a partial answer to the moral design problem. Consider,

> *The Restricted Pareto Principle** (RPP*)*:* In collisions where (i) harm to at least one person is unavoidable, and (ii) a choice about how to allocate harm between different persons is required, then *if* there exists a unique Pareto efficient allocation of harm across different persons, *then* other things being equal, programming a driverless car to bring about the Pareto efficient allocation of harm is justified.

[7] The jury found that "there was no appreciable chance of saving life except by killing someone for the others to eat" (*Dudley and Stephens* 1884: 275).

[8] It might be objected that the Restricted Pareto Principle (RPP) does not provide a complete account of the conditions under which the doctrine of necessity provides a justification for harm. This might be true, at least insofar as RPP does not consider the normative positions of the relevant parties. However, my aim is to vindicate the narrow reading of Santoni de Sio's legal-philosophical approach. That is, I want to use the law as a starting point for theorising about the moral design problem. It strikes me that RPP is a plausible decision-theoretic reading of Brooke LJ's distinction between the facts of *Re A* and *Dudley and Stephens*, and that it can help us formulate a partial answer to the moral design problem which three of our best moral theories can agree upon.

I shall now argue that RPP* is a decision-rule which at least three of our best moral theories can agree on as a partial answer to our question.

5 Overcoming the Problem of Disagreement

I now argue that, in the scenarios where it applies, RPP* is a principle which utilitarians, contractualists and deontologists have reason to accept. Their reasons for accepting RPP* are grounded in the moral principles which they endorse.

I start with utilitarianism. I take it that most utilitarians believe that, in collisions where harm is unavoidable, driverless cars should be programmed to bring about the outcome which maximises utility (Bonnefon et al. 2016). In the collisions where it applies, RPP* justifies programming a driverless car to bring about an outcome only if that outcome is utility-maximising. This is because if one outcome Pareto-dominates another, then the dominating outcome has strictly greater utility than the dominated outcome (Coleman 1980: 515). As a unique Pareto efficient allocation of harm Pareto-dominates all other outcomes, it follows that the unique Pareto efficient allocation of harm is utility-maximising. So, the utilitarian has reason to accept RPP* because, on their view, driverless cars ought to be programmed to maximise utility in collisions, and RPP* does this in all the collisions where it applies.[9]

I now consider contractualism. I discuss T.M. Scanlon's (1998) account of contractualism, as this is plausibly the most sophisticated articulation of the view. Scanlon argues that the rightness of moral principles is determined by the justifiability of those principles to individuals affected by their prescriptions. An act is permissible, Scanlon argues, if it is permitted by any principle for the general regulation of behaviour that no one could reasonably reject. Roughly, an individual has grounds to reasonably reject a principle when it gives insufficient weight to her moral claims. These claims include considerations about her utility, her being treated unfairly, her not being accorded appropriate respect, and so on. What matters is that the considerations are *personal* to the individual in question, in the sense that the considerations have to be about how a proposed principle would treat her or would imply about her.

The contractualist has at least three reasons to accept RPP* as a partial answer to the moral design problem. First, an important feature of Scanlon's view is the *individualist restriction*, which holds that we cannot aggregate the moral claims of separate persons. The claims of each person must be assessed in isolation. RPP* determines that an outcome in a collision is justified by considering the utility of each affected person in isolation. For each person, a unique Pareto efficient outcome is such that either (i) the individual would rationally prefer that outcome to any other, as it causes the least harm to *them*; or (ii) the individual would not strictly prefer any alternative outcome, as no outcome will bring about strictly less harm to *them*. The aggregation of moral claims is not relevant to determining the unique Pareto efficient outcome, and in turn, it is not relevant to RPP*. Hence, RPP* is consistent with the individualist restriction.

[9] It might be objected that utilitarians would also consider the long-term consequences of accepting a rule like RPP*. Perhaps, if the public knew that under some conditions, driverless cars are programmed to kill their passengers, they would be reluctant to purchase driverless cars. As driverless cars are much safer than non-driverless cars, perhaps adopting a rule like RPP* would, therefore, cause more deaths in the long run. I am unconvinced. RPP* justifies programming the car to kill its passenger only if the passenger was going to die anyway. I doubt people would be reluctant to purchase driverless cars in light of this.

Second, RPP* can be understood as a welfarist principle, insofar as it considers only the *utility* of affected persons conditional on each outcome. Scanlon admits considerations about utility as relevant to whether a principle can reasonably be rejected. But utility is not the only consideration. It might therefore be objected that RPP* is non-contractualist, because it fails to account for non-welfarist considerations like fairness and responsibility. I find this objection unconvincing. In explaining why, I hope to show that RPP* captures another feature of Scanlon's view.

Scanlon argues that:

> In many cases, gains and losses in well-being (relief from suffering, for example) are clearly the most relevant factors determining whether a principle could or could not be reasonably rejected (Scanlon 1998: 215).

I imagine that driverless car collisions where harm is unavoidable and there exists a unique Pareto efficient allocation of harm are cases where the welfare of each person is what matters most to the justifiability of the moral principles in question. At the very least, it strikes me that fairness and responsibility are not obviously relevant to whether RPP* can reasonably be rejected.

Scanlon (1998: 212–213) maintains that a person can reasonably reject a principle on the grounds of fairness only if the principle arbitrarily favours one person over another. RPP* justifies saving one person and killing another only if the individual killed would die irrespective of the driverless car's actions. In this respect, the decision to kill one person to save another is non-arbitrary, and RPP* cannot reasonably be rejected on the grounds of fairness. As for responsibility, I take it that *even if* a person is responsible for causing a collision, they can reasonably reject any principle which mandates that the driverless car should bring about her death, when this could be prevented at no cost (in welfare terms) to any other affected party.[10] So, it strikes me that a contractualist response to the moral design problem would take welfare as the principal moral consideration; and to this extent, RPP* captures another feature of the contractualist position.

Third, Scanlon endorses a principle which is stronger than RPP*. According to

> *The Rescue Principle:* […] if you are presented with a situation in which you can prevent something very bad from happening, or alleviate someone's dire plight, by making a slight (or even moderate) sacrifice, then it would be wrong not to do so (Scanlon 1998: 224).

If welfare is the only relevant consideration in driverless car collisions where harm is unavoidable, then RPP* never requires an affected party in a collision to make a sacrifice. All parties would be at least as badly off on any other alternative. So, if the Rescue Principle is true, then programming a driverless car to bring about the same degree of harm to someone in a different way in order to save another affected party from death or serious harm is not something that can reasonably be rejected.

[10] It might be objected that RPP* can reasonably be rejected on *broader* grounds of responsibility. I have not considered the *incentives* that RPP* would create. Perhaps if RPP* were implemented in driverless cars, then reckless drivers in manual vehicles would be assured that driverless cars would act to save them from death, if this could be done without causing additional harm to other parties in the collision. So, plausibly most road-users (including driverless car passengers) could reasonably reject RPP* on the grounds that it incentivises irresponsible or reckless driving. I cannot dispel this objection in its entirely. However, there is a slim probability of a *particular* reckless driver causing a collision with a driverless car, where this collision has a unique Pareto efficient allocation of harm *in the driver's favour*. So, we can at least say that reckless drivers do not have *good reason* to believe that RPP* would significantly improve their survival prospects should (i) RPP* be implemented into driverless cars and (ii) they continue to drive recklessly.

This concludes the contractualist argument. I now turn to the deontological case. There are a family of deontological considerations, and there is not scope to address all of them. I consider two of the most important: the distinction between killing and letting-die and the problem of incommensurability. I argue that RPP* respects both. I use an example collision scenario to aid the discussion. Consider,

> *Head on Collision:* A driverless car is travelling at a safe speed along a narrow country road. It approaches a blind bend. As the driverless car comes around the bend, it detects a human-driven car travelling at high speed in the opposite direction. If the driverless car brakes, the other car will crash into it, most likely causing the death of its passenger and the driver in the other vehicle. It can, however, swerve off the road, most likely causing the death of its passenger, but saving the life of the driver in the other vehicle.

Here, the unique Pareto efficient allocation of harm is where the passenger in the driverless car dies, and the driver in the other vehicle survives. RPP* holds that programming the driverless car to swerve and cause the death of its passenger is justified.

Does RPP* respect the killing and letting-die distinction? Minimally, it is true that RPP* does not take as morally salient the distinction between programming the car to *act* and programming the car to *omit to act*. But to infer from this that RPP* does not respect the killing and letting-die distinction, insofar as it permits *killing* the passenger to avoid *letting both parties die*, relies on a particularly strong conception of the killing and letting-die distinction. Francis Kamm (2007) offers a plausible account of the distinction which RPP* does satisfy. Here is Kamm's account:

> (1) In killing, we introduce a threat that was not previously present; in letting die we do not interfere with a currently present threat. (2) In killing, we act; in letting die, we fail to act. Presumably, the nonaction is not a mere omission but a refraining. For example, we are not asleep while someone dies, but we consciously choose not to aid. (3) (i) In killing, we cause someone to lose a life that he would have had independently of our efforts at that time; (ii) in letting die, someone loses a life that he would have had only with our help at that time. (4) In killing, we initiate an interference with the victim; in letting die, we avoid being interfered with (by having to aid) (Kamm 2007: 18).

In Head-On Collision, at least condition (3.i) is not satisfied if the driverless car is programmed to swerve and cause its passenger's death. It is not the case that in programming the car to swerve, we cause the passenger to lose a life that she would have had independently of our programming the car to swerve. RPP* also does not justify letting the passenger die. Condition (3.ii) is not satisfied as there is no way to save the passenger from death. These considerations can be generalised. RPP* justifies harming someone *only if* they would receive at least the same harm on any alternative. RPP* is therefore neutral with respect to Kamm's account of the killing and letting-die distinction. It never justifies *killing* anyone, and so, the distinction is respected.

RPP* also respects the problem of incommensurability insofar as it does not require the programmers of driverless cars to weigh-up the comparative value of the lives of parties involved in collisions. RPP* is based on Brooke LJ's distinction between *Re A* and *Dudley and Stephens*. Like Brooke LJ's conditions for necessity in *Re A*, RPP* does not require the driverless car's programmer to choose which person to save. For all collisions with a unique Pareto efficient allocation of harm, no such choice is required. So, RPP* is at least consistent with the view that we should not, or cannot, make comparisons about the relative values of different lives.

I have argued that RPP* can be defended using three different moral perspectives: utilitarianism, contractualism and deontology. The arguments here are not conclusive. But I hope, at least, to have illustrated that the following claim is plausible: RPP* is a partial answer to the moral design problem which many of us can accept *despite* our disagreements about which moral principles offer the correct account of justified harm.

6 Conclusion

In this paper, I developed Santoni de Sio's (Santoni de Sio 2017) legal-philosophical approach to the moral design problem. In §2, I distinguished two readings of the approach. On the wide reading, the aim is to use the law to resolve our disagreements about the correct moral account of justified harm. On the narrow reading, the aim is to use the law to bring us to a consensus on the moral design problem *despite* our disagreements about the correct account of justified harm. In §3, I explained Santoni de Sio's answer to the moral design problem, and argued that, on either reading of the legal-philosophical approach, his answer falls short of the aim of the approach. In §4 and §5, I salvaged the narrow reading, to an extent. I articulated an alternative reading of the doctrine of necessity based on a restricted Pareto principle. I used this principle to formulate a partial answer to the moral design problem, which provides a justification for harm in some, but not all, of the kinds of collisions with which the moral design problem is concerned. I then argued that utilitarians, contractualists and deontologists have reason to accept the restricted Pareto principle. If my arguments are successful, then the restricted Pareto principle allows us to reach a partial consensus on the moral design problem despite our disagreements about which principles provide the correct account of justified harm.

I have not discussed one further advantage of the restricted Pareto principle over Santoni de Sio's answer to the moral design problem. Santoni de Sio's answer is qualitative. The restricted Pareto principle is formulated in decision-theoretic terms, such that it could, in principle, be programmed into a driverless car. There is also scope to adapt the restricted Pareto principle to account for collisions where the driverless car has incomplete information about the harms to each affected person conditional on each alternative. I suspect that similar moral arguments could be constructed in favour of this modified principle.

Acknowledgements I am extremely grateful to Brad Hooker and Richard Pettigrew for their extensive comments on earlier drafts and for several thought-provoking discussions. I also owe a great deal of thanks to Farbod Akhlaghi-Ghaffarokh, Michael Hauskeller and Niall Paterson for many helpful discussions and comments. I am also indebted to two anonymous reviewers whose comments helped me to improve the paper significantly. This research was funded by an Arts and Humanities Research Council PhD Studentship, awarded through the South, West and Wales Doctoral Training Partnership.

References

"SAE Information Report (J3016) Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems." (n.d.)

Arnolds EB, Garland NF (1974) The defense of necessity in criminal law: the right to choose the lesser evil. J Crim Law Criminol 65(3):289–301

Austin JL (1956) A Plea for excuses. In: Austin JL (1961) Philosophical papers, edited by J. O. Urmson and G. J. Warnock. Oxford University Press, Oxford

Bohlander M (2006) Of shipwrecked sailors, unborn children, conjoined twins and hijacked airplanes—taking human life and the defence of necessity. Journal Crim Law 70(2):147–161

Bonnefon J-F, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. Science 352(6293): 1573–1576

Christie GC (1999) The defense of necessity considered from the legal and moral points of view. Duke Law J 48(5):975–1042

Coleman JL (1980) Efficiency, utility, and wealth maximization. Hofstra Law Review 8(3):509–551

Gerdes JC, Thornton SM (2016) Implementable ethics for autonomous vehicles. In: Maurer M, Gerdes JC, Lenz B, Winner H (eds) In *Autonomous Driving: Technical, Legal and Social Aspects*. Springer, Berlin Heidelberg, pp 87–102

Hart HLA (1961) The concept of law. Clarendon, Oxford

Kamm F (2007) Intricate ethics: rights, responsibilities, and permissible harm. OUP, Oxford

Leben D (2017) A Rawlsian algorithm for autonomous vehicles. Ethics Inf Technol 19(2):107–115

Lin P (2016) Why ethics matters for autonomous cars. In: Maurer M, Gerdes JC, Lenz B, Winner H (eds) *Autonomous Driving: Technical, Legal and Social Aspects*. Springer, Berlin Heidelberg, pp 69–85

Litman, Todd. Autonomous vehicle implementation predictions. Victoria Transport Policy Institute, 2014

Mouse's Case (1608) Michaelmas term 6, JMS I vol 12

R v Dudley and Stephens (1884) QBD 273

Re A (Conjoined twins) (2001) 2 WLR 480 (CA)

Reninger v Fagossa (1551) 1 Plowd. 1, 75 Eng. Rep. 1

Santoni de Sio F (2017) Killing by autonomous vehicles and the legal doctrine of necessity. Ethical Theory Moral Pract 20(2):411–429

Scanlon TM (1998) What we owe to each other. Harvard University Press, Cambridge