Delivery Note

Bavarian state library, Munich
Ludwigstr. 16
D-80539 Muenchen

Fon: ++49-89-28638-2451
Fax: ++49-89-280-9284
E-mail: doklief@bsb-muenchen.de

Delivery address

Dokumentlieferdienste \50z.Hd. Frau Wettstein\51
Universitaet Konstanz - KIM / Bibliotheksdienste
Universitaetsstr. 10
D-78464 Konstanz

Data concerning order:

Order date:	2024-02-15 14:39:01
Order number:	SUBITO:VE24021500508 /
Customer name:	Universitaet Konstanz - KIM / Bibliotheksdienste
User account:	HSL9700002
Delivery date:	2024-02-15 18:22:04
Delivery priority:	NORMAL
Delivery way:	Email
E-mail address:	sharon.baute@uni-konstanz.de

Remarks concerning delivery:

Data concerning document:

Signature:	2020.46375
Author:	S. Matthew Liao
Title:	Ethics of artificial intelligence
Year:	
Volume / Issue:	
Pages:	uknown
Author of article:	F. M. Kamm
Title of article:	The Use and Abuse of the Trolley Problem: Self-Driving Cars, Medical Treatments, and th
ISSN:	
ISBN:	9780190905033
CODEN:	

Your comment concerning the order:

subito copyright regulations

Copies of articles ordered through subito and utilized by the users are subject to copyright regulations. By registering with subito, the user commits to observing these regulations, most notably that the copies are for personal use only and not to be disclosed to third parties. They may not be used for resale, reprinting, systematic distribution, emailing, web hosting, including institutional repositories/archives or for any other commercial purpose without the permission of the publisher.

Should delivery be made by e-mail or FTP the copy may only be printed once, and the file must be permanently deleted afterwards.

The copy has to bear a watermark featuring a copyright notice. The watermark applied by subito e.V. must not be removed.

FTP

Bestelldatum: **2024-02-15 14:39:01**

Bayerische StaatsBibliothek

NORMAL

Kopie

SUBITO-VE24021500508

Universitaet Konstanz - KIM / Bibliotheksdienste
Dokumentlieferdienste (z.Hd. Frau Wettstein)
Ms Jana Wettstein
Universitaetsstr. 10
78464 Konstanz

Ben.-Gruppe: USER-GROUP-1
Tel: +49 7531 882824
Mail: docdel@subito-doc.de

Fax:

Subito-Kundennummer:
HSL9700002
Subito-Bestellnummer:
SUBITO-VE24021500508

2020.46375

Jahrgang:

Band/Heft:

Seiten: uknown

Verfasser: F. M. Kamm

Titel: The Use and Abuse of the Trolley Problem: Self-Driving Cars, Medical Treatments, and the Distributio

Ethics of artificial intelligence
ISSN:

Bemerkung:

Beschreibung: If possible, please also include the introduction of the book

Die Abrechnung dieser Lieferung erfolgt über die subito-Zentralregulierung

Bei Rückfragen wenden Sie sich bitte innerhalb von 10 Tagen an die Bayerische Staatsbibliothek, Direktlieferdienste
Tel. ++49 89 28 638-24 51, dokumentlieferung@bsb-muenchen.de

2

The Use and Abuse of the Trolley Problem

Self-Driving Cars, Medical Treatments, and the Distribution of Harm

F. M. Kamm

In this chapter I first briefly present cases that are standardly considered "Trolley Problem" Cases along with standard moral judgments about permissible conduct in these cases. Next, I consider the ways in which many standard car driving cases differ as a conceptual matter from standard Trolley Problem Cases with which some compare them. I argue that the cases involving cars raise distinctive moral issues different from the distinctive issues raised by standard Trolley Problem Cases. I also consider how some medical cases differ from some standard trolley cases with which some compare them. Finally, I discuss some moral issues raised by self-driving cars by comparison to Trolley Problem Cases, including the role of those who would program the cars and the liability to harm of pedestrians, drivers, and passengers.

2.1. A Hypothetical Case

I once considered what I called the Ambulance Case.[1] In it society was to decide ex ante (i.e., in advance of knowing who would be affected one way or another) how an ambulance should be programmed when it came to a choice between saving people by rushing them to the hospital and harming pedestrians on the route or letting the patients die but harming no pedestrians. I imagined that the ambulance could be made to detect how many people it was carrying and how many pedestrians would be harmed, and, to simplify matters, I assumed (as I do here) that the life of each person was at stake and that they were alike in morally relevant respects. One question I considered was whether we should deliberately program our ambulance carrying five people to continue on its route by having the program disable a stopping mechanism whenever more lives would be lost by its stopping than by running over one person on the road. I argued that even though ex ante (at the time we decide on how to program the ambulance) each person in the society would maximize his chances of survival by not allowing the

F.M. Kamm, *The Use and Abuse of the Trolley Problem* In: *Ethics of Artificial Intelligence*. Edited by: S. Matthew Liao, Oxford University Press (2020). © Oxford University Press. DOI: 10.1093/oso/9780190905040.003.0003

ambulance to stop, it could be wrong to program the ambulance in this way, just as it would be wrong for a driver in control of the ambulance to drive over the one person to get the five to the hospital. Certainly, the fact that only a program, not a person, would disable the stopping mechanism at the time of impact does not remove responsibility for this happening from the people who programmed the ambulance.

2.2. Standard Trolley Cases

Recently, reality has caught up with hypothetical cases like the Ambulance Case in which vehicles can be programmed to move in various ways. One such instance is the design of so-called self-driving cars of which there could be at least two types: (1) those that have no person at all driving them and that operate completely on the program designed for them (call this the Complete Case) and (2) those that have a person driving them but whose program can override or supplement a driver's control at crucial points (e.g., the car will stop despite the driver trying to continue; call this the Partial Case).[2]

Some have thought that what is known as the Trolley Problem, a topic in normative ethical theory, might help us with the practical problem of creating programs for self-driving cars. Here is a description of the basic Trolley Case created by Philippa Foot:[3] A driver is on his out-of-control trolley that is headed toward killing five innocent people on a track. To save them he can only turn the trolley to a sidetrack where, he foresees, the trolley will instead kill one other person standing on that track. A variation on this case introduced by Judith Thomson involves a bystander, not the driver, deciding whether to turn the trolley (with the same effect) when the driver cannot do so (the Bystander Case).[4] It is commonly thought that since the driver started the trolley, for him it is a choice of killing five or killing one. However it might be argued that if, independently of any act or omission of his, the trolley goes out of control, how he comes to cause the death of the five would be very different from his deliberately turning the trolley when he foresees it will kill someone else. For the bystander it is a choice of letting five die or killing one.[5] In none of these cases is the person who turns the trolley or who is on the trolley at risk of dying. In none of these cases is the person who would be killed by the redirected trolley the cause of its threatening the five or obligated to either give his life or assume a position in which he will be killed as a side effect in order to save the five other people. He would also lose his life by someone else imposing the loss on him, not by his imposing it on himself.

Many have thought that in these cases the driver is morally required and the bystander is at least permitted to do what will minimize the number of people

who would be killed by redirecting the trolley and that there is some principle that explains the obligation and permissibility, respectively. (These cases involve a choice between five or one being killed, but the argument that would justify killing one to save five arguably would also justify redirecting toward one person to save even two other people and redirecting toward any smaller number of people to save a larger number of other people. Indeed I think that if an argument could not justify killing one to save two from the trolley, it could not justify killing one to save five from it.)

By contrast, many think it would be impermissible for either the driver or the bystander to do what is necessary to save the five people by killing one other person in the following ways: (1) topple someone from a bridge in front of the trolley so that his body stops it and the trolley kills him (Topple Case)[6] or (2) use a small bomb whose explosion would stop the trolley from hitting the five, though a piece of the bomb would fly off, killing an innocent pedestrian (Bomb Case). It is thought that some principle explains why these things are impermissible even though they would also minimize the number of innocent people killed. We could think of the Trolley Problem as explaining why killing is permissible in the Trolley and Bystander Cases but not in the Topple and Bomb Cases. This problem is part of the more general debate between act consequentialists, who think it always permissible to bring about the best consequences, and nonconsequentialists, who deny this.

2.3. Other Cases

I have described what philosophers who have discussed the Trolley Problem take to be some standard cases involved in it. Let us now consider how these cases compare conceptually with those that have been described as Trolley Problem Cases by others, including contributors to popular internet sites who aim to acquaint the general public with the problem in connection with ordinary and self-driving cars. At this stage, I am concerned only with conceptual similarities and differences and put to one side moral judgments of the cases. Nevertheless I am concerned with similarities and differences because they may be morally significant. For one example, Chris Rampolla (who describes himself as "a philosopher by training") says the following in his online article:

> Suppose that you're driving your car in the right hand lane of a one-way street on a winter evening. As you approach a red light at an intersection, you tap the brakes and begin to skid. Ahead of you the left lane is closed and is blocked by a concrete barrier in front of a crosswalk. There are no obstructions in the right lane. A pedestrian has legally entered the crosswalk on the right side of

the street and is attempting to cross over to the other side. You have just enough time and just enough control of the car to make a decision about which lane to enter but you cannot stop your car. Should you choose to continue on in the right lane, the pedestrian will be struck by the car and will likely die. Should you choose to direct your car into the left lane, the collision of your car with the barrier will save the life of the pedestrian but will very likely kill you, the driver. What do you choose to do? Now ask yourself that same question, except this time consider that your child is in the car and would likely die from an impact with the barrier. Next consider that the pedestrian in the road is also accompanied by a child. Still further, consider that this time your spouse and child are in your car with you and there are three elderly people in the crosswalk. Has your choice changed? More generally, what is the moral thing to do in each of these situations and is there any commonality between them? These are modern versions of a philosophical problem known as *The Trolley Problem.*[7]

2.3.1. Threats and Nonthreats

Rampolla here describes the Trolley Problem as being instantiated in a case where a driver of a car that cannot be stopped has to decide whether to let the car continue on to (very likely) kill a pedestrian, or turn it and (very likely) kill himself.[8] (Henceforth, to simplify, I will assume someone will definitely be killed, as is assumed in the standard trolley cases.)[9] But the standard Trolley Problem Cases discussed by philosophers are about whether one may kill some innocent, nonthreatening person so as to either not kill or save other innocent, nonthreatening people from a threat already facing them. These cases do not, as in Rampolla's case, involve someone who is presenting a threat of death to others (the driver) deciding whether to kill innocent, nonthreatening people rather than kill himself. Philosophers might refer to the latter sort of case as an "innocent threat case." Such cases are distinguished by the threatener himself, a bystander, or a potential victim having to decide whether to harm the threatener, who became such despite being innocent of any moral wrongdoing, or allow him to proceed to harm some nonthreatening people.

I am not concerned with legislating the use of the terms "Trolley Problem" and "Innocent Threat Case." I am concerned with the possibility that the difference I have identified between standard trolley cases and the one in Rampolla's example may be morally significant and that it is wrong to assume without additional argument that a driver who threatens people should weigh his life or have it weighed by others as the lives of innocent, nonthreatening persons should be weighed when one decides who will be harmed. To keep this moral issue in mind, I shall distinguish the Trolley and Innocent Threat Cases in the ways I have just described (though others may follow a different practice).[10]

There are different types of Innocent Threat Cases which make it clearer why the threatener is considered morally innocent. Robert Nozick described a man unwillingly shot out of a cannon, headed toward landing safely on someone who, however, would be killed by the impact.[11] May the potential victim or a bystander kill the innocent threat to prevent his fall from killing the victim? In other Innocent Threat Cases (sometimes called cases of minimal responsibility for harm)[12] the threatener is not totally inactive in causing the threat but is not at fault in his actions (e.g., a psychotic killer, a nonnegligent driver whose brakes fail in his well-maintained car). All these cases involve unjustified threats, but might there be other Innocent Threat Cases where a nondeliberate threat is justified? For example, consider a pilot fighting on the just side in a war who unintentionally bombs a military facility that it would have been permissible to intentionally bomb given that there is only permissible collateral harm to civilians. In standard just war theory anyone who presents a threat is considered a "noninnocent," but morally speaking one might consider him innocent. May these threatened civilians nevertheless shoot him down to protect themselves (especially if this will not affect the good military effect of his bombing)? In this case it is the pilot who is the threat, for if we could eliminate him and thus his control of the bombs, the threat to civilians would stop. However, in cases involving an already out-of-control car it is strictly the vehicle that is the threat; getting rid of the driver may not stop the threat,[13] though doing something to the vehicle that as a side effect will kill the driver will stop the threat. Nevertheless I think in an extended sense, the driver in such cases is an innocent threat if he was driving the vehicle before he lost control of it.

In standard Innocent Threat Cases, the choice is between the original potential victim(s) being harmed and the threatener being harmed (whether the harm to him would be imposed by the threatener himself or by someone else). Suppose a third option is added so that the threat can be redirected to another nonthreatening victim. Then we could possibly have a combined Innocent Threat and Trolley Problem Case as I am using these terms. However, if a driver had to bear the burden of being harmed rather than his original potential victims, it is unlikely to be permissible for anyone except the new potential victim to decide that the threat should be turned to her, holding losses to all constant. Then problems distinctive to standard Trolley Problem Cases would not arise.

2.3.2. Imposing Harm versus Having It Imposed

Note also that in the standard Trolley Cases, unlike the case I cite from Rampolla, the person who decides what to do with the trolley (either the driver or a bystander) is not one of those who might be harmed. Hence those who stand to be

harmed in standard Trolley Cases have harm imposed on them by others rather than imposing it on themselves. By contrast, in Rampolla's case the driver faces the option of harming himself.[14] One should not assume without argument that what may be imposed on one is the same as what one has a duty to impose on oneself.

Hence, unlike the standard Trolley Cases, Rampolla's case involves both the possibility of harm to the threatening driver and the driver having to decide whether to impose the harm on himself. The first factor may make it easier to justify harm to one person; the second factor may make it harder to demand it. Suppose one could show that the driver was permitted to treat himself like any nonthreatening possible victim. The fact that no such victim would have to redirect the threat away from others to himself would imply that the driver (who is actually the threat) would not have to redirect in a way that kills himself.[15]

2.3.3. Passengers and Varied Characteristics

As the quote from Rampolla shows, he also considers cases in which there are people in the threatening vehicle who are simply passengers and who stand to be harmed if the threatening driver himself decides to avoid harming pedestrians. Cases with passengers are unlike the standard Trolley Cases in which there is no one on the trolley who can be harmed, and it is also a complication of the standard Innocent Threat Case. If there are passengers in a threatening vehicle, it would be another mistake to assume without argument that they are to be treated as morally equivalent to pedestrians who would be hit by the threat and that their lives should be weighed in the same way as pedestrians.

Some of Rampolla's cases also assign different qualities to the different people who might be harmed (e.g., some are old, some young, some are strangers, some are one's spouse or child).[16] By contrast the standard Trolley and Innocent Threat Cases abstract from such distinctions in order to determine (1) whether in the Trolley Cases differences in how we come to kill people (e.g., by redirection, by toppling, etc.) make a moral difference and (2) whether in the Innocent Threat Cases the difference between being a threat (albeit morally innocent) rather than a pedestrian makes one more liable to be harmed. As in science, it has been thought in philosophy that to test for the moral significance of a factor, we should insofar as possible hold other variables constant. Nevertheless, varying additional characteristics to see how they affect one's decisions about sparer cases need not be a methodological mistake since otherwise crucial factors can be overridden or have their effect altered in changing contexts.

2.3.4. Possible Principles

A final point about Rampolla's discussion is that he describes Foot as supporting a revised version of the Doctrine of Double Effect (DDE) to justify the driver turning the trolley. He says:

> A modern interpretation of the doctrine of double effect was put forth by Philippa Foot in 1967. The problem (at the time) had nothing to do with driving, but instead was one of a number of thought experiments she used to examine the morality of abortion. As an example of *double effect* [emphasis added] she suggested the following:
>
> "The steering driver faces a conflict of negative duties since it is his duty to avoid injuring five men and also his duty to avoid injuring one. In the circumstances he is not able to avoid both, and it seems clear that he should do the least injury he can. The judge, however, is weighing the duty of not inflicting injury against the duty of bringing aid. He wants to rescue the innocent people threatened with death but can do so only by inflicting injury himself. Since one does not in general have the same duty to help people as to refrain from injuring them, it is not possible to argue to a conclusion about what he should do from the steering driver case."

But it is not true that the distinction Foot draws in the part of her article quoted by Rampolla is a modern interpretation of the DDE. The DDE distinguishes morally between harm that happens as a side effect and harm that is intended. It claims the latter is impermissible even as a means to a greater good while the former at least does not rule out pursuing a greater good. Foot proposes an alternative to the DDE that has nothing to do with whether one intends harm (or harm is a means). It focuses on a moral distinction between harming and not aiding even when harm is merely foreseen as a side effect and neither intended nor causally a means. For example, Foot thought that using a gas in surgery to save five people from their illness is ruled out if this gas will also cause someone else's death as a mere foreseen side effect. (I will call this the Gas Case.) (Note that her claim that the duty not to harm is stronger than the duty to aid also might rule out turning the trolley in the Bystander Case since the bystander would harm one person to aid five.)[17]

2.4. New and Old Threats?

Other examples of the problematic use of the Trolley Problem occur in medical ethics discussions. One instance is in the work of Dr. Marya Zilberberg.[18]

She correctly identifies a version of the Trolley Problem as reconciling the apparent permissibility of the driver diverting the trolley and the impermissibility of a bystander toppling a person in front of the trolley to stop it even though in both cases five people would be saved from death and one person would be killed. She goes on to compare these cases with the use of mammography which is said to save eight women who would otherwise have died from cancer for every thousand mammogrammed but lead to the death of at least one woman who would not otherwise have died (due to false positives). She says about this Mammography Case:

> Well, then we have the trolley problem, don't we? We are potentially sacrificing 1 individual to save 8. And who does the sacrificing is where the variations of the trolley problem come in. . . . The payer certainly sees this issue as the original formulation of the problem: Why not throw this financial switch to achieve net life savings? But for a clinician who deals with the individual patient this may be akin to pushing her over the bridge toward a potentially fatal event.

The first thing to note is that unlike the driver in the "original formulation of the problem," the payer did not have a role in causing the threat that must be dealt with (in this case, cancer). Nor is he diverting the cancer that threatens the life of some women toward fewer women. The payer is helping to pay for a new means to help some women (the mammogram) which presents a bad side effect threat to a smaller number of other women. (This is analogous to using the small bomb to stop the trolley when the bomb would kill another person.) This also implies that the doctor who orders mammograms that as a side effect harm a woman is not harming a woman as a *means* to help save eight others, which would be akin to toppling someone in front of the trolley as a means of stopping it. This is shown by the fact that if mammograms the doctor orders did not cause the harm to one woman as a side effect, this would not reduce mammograms' effectiveness in saving eight out of a thousand women; the threat and harm to the one is not needed to save the eight.

An additional aspect of the Mammography Case to which Dr. Zilberberg points is that ex ante each person could be either one of the eight who will be benefited or the one who will be harmed. Hence she thinks it is important to understand the patient's attitude to risk. And indeed introducing a means that will unavoidably risk harm as a side effect to each of those who, as far as we know ex ante, also get from that means a greater chance to benefit may be permissible. However, this does not imply that it is permissible to use means that will help others when it is known that the means will as a side effect impose certain death on a particular other person, and it is still possible to prevent its doing so by not using the means at the time we know it would present the threat (as in

the Bomb and Gas Cases). This is so even if it is permissible in the Trolley and Bystander Cases to impose certain death on a particular other person when one could avoid it (independent of any ex ante calculation of risks and benefits). As already noted, it is one aspect of the Trolley Problem to explain these differences in permissibility.

Not distinguishing between the Trolley Case and cases like Bomb and Gas is also exemplified by some discussions in medical ethics that try to analogize the use of electronic cigarettes to diverting the trolley.[19] That is, many people will die of smoking. Suppose we can reduce their numbers by converting them to use of electronic cigarettes, which are somewhat less bad for them. However, suppose also that as a side effect of this policy a smaller number of other adults who would not have smoked will also take up electronic cigarettes, thus becoming worse off than they would otherwise have been. (This is a hypothetical case insofar as it abstracts from real effects, especially on underage users.) This case is not analogous to diverting the trolley away from more people to fewer people for the same reason that the Bomb and Gas Cases are not like diverting the trolley. In this case there is a prior threat (of cigarettes) to many people. If we introduced e-cigarettes as a means to help them, we would not be diverting cigarette use but introducing a new means that would have the side effect of harming (by hypothesis) a smaller number of people who never smoked. We may wonder whether this conceptual difference makes a moral difference, but it *is* a conceptual difference. It is just important to remember that not all cases that involve only foreseeably killing fewer nonthreatening people to save a greater number of other nonthreatening people are like the standard Trolley Case in which diverting a threat seems permissible.

However, it is worth noting that, as described, the Electronic Cigarette Case differs from the Gas and Bomb Cases in at least two significant ways: (1) The e-cigarettes that reduce deaths among cigarette smokers do not directly cause harm to others, as do the bomb and the gas. It is only because an intervening agent takes up smoking e-cigarettes that he may be harmed. Helping the cigarette smokers by getting them to use e-cigarettes at most *enables* the harm to another new group of e-smokers by making available to them the new option of e-cigarettes. (2) The likelihood of death to each of the new e-cigarette smokers is (assumed) less than the likelihood of death to the original cigarette smokers. (This is so even if there are eventually more deaths due to e-cigarettes than to cigarettes because more rather than the hypothesized fewer new smokers will use them.) So unlike the Bomb and Gas Cases, the probability of death occurring to each of the newly threatened people is less than the probability of death occurring to each of the people originally threatened by cigarettes. These two factors may make the grounds that rule out using the bomb and gas inadequate to rule out the introduction of e-cigarettes in the hypothesized circumstances despite their bad side effects.

Here is an implication for self-driving cars of what we have just said about the Mammography and Electronic Cigarette Cases: The permissibility of killing some to save others in the standard Trolley Case is relevant to programming cars only if programming is about redirecting the threat that the car itself presents. However, this does not mean that to be morally like the Trolley Case anyone who dies as a result of a threatening car being redirected must also be killed by that very car. For example, suppose that the turning trolley (or car) caused a new threat of a rockslide that killed a pedestrian. Turning the trolley (or car) is still a permissible solution in the trolley-type case because it is the trolley (or car) turning away from the five people that causes the new threat,. This is not so with a newly introduced means to turn the trolley like the bomb which presents a new threat.[20]

The overall conclusion of considering these discussions that try to make use of the Trolley Problem (in both sections 2.2 and 2.3) is that they often fail to recognize the very distinctions the Trolley Problem Cases are about. These distinctions may or may not be morally significant, but not recognizing them is itself problematic.

2.5. Particular Moral Issues in Self-Driving Cars

Though some would disagree, let's assume for the sake of argument that all the judgments I cited in section 2.2 about what is commonly thought to be permissible, obligatory, and prohibited in the standard Trolley Problem Cases are correct and that there are principles that justify these judgments.[21] This would imply that we know what morally should or may be done in many cases. If we do not have access to the principles that underlie our judgments, we cannot program the principles into cars, though we might provide rules for what they should do in a variety of cases.[22] However, there may be other ways in which cars can "learn" what to do besides being programmed with rules or principles. It is said that, like people, machines exposed to various situations can self-learn to make correct choices without following explicit rules or interpreting known principles. These are complex issues about learning that I will not discuss here. For simplicity I will refer to the cars in Complete and Partial Cases as being programmed with principles that lead them to behave properly.[23] My concern here is not *how* to make self-driving cars behave properly but *what* the proper way for them to behave is and whether there are substantive moral (rather than merely conceptual) differences between what is morally proper in standard Trolley Cases and in the Complete and Partial Cases.

Here are some issues to consider.[24]

2.5.1. Why Have Self-Driving Cars?

The primary benefit of Complete or Partial cars is that they will prevent deaths by preventing situations in which any person's life is threatened. Their primary benefit is not to merely reduce lives lost once something has already gone wrong and at least someone will have to die.

In the Trolley Cases something has already gone wrong since the trolley has gone out of the driver's control with respect to the five, and the question is who is to die when someone must die. Hence the primary issue with which the Trolley Problem (and also Innocent Threat Cases) deal—what to do when someone must die—would arise only when self-driving cars have failed to satisfy the primary reason for having them. If self-driving cars got into dangerous situations more often than cars completely under a driver's control, the fact that they could be programmed to do a better job than a driver of minimizing the harm they would cause would not speak as strongly in their favor. However, if these cars got into life-threatening situations less often than human-driven cars, that they were worse in determining who will die when someone must might not speak very strongly against them.[25]

2.5.2. Do Moral Principles That Apply to Persons Apply to Machines?

It may be said by some that the sorts of moral prohibitions that make it impermissible, for example, to kill one person to save five in the Topple and Bomb Cases, are relevant to the conduct of persons but not to machines because the prohibitions are grounded in "the agent's personal point of view" of acting in certain ways and automated cars are not agents that have a personal point of view. This so-called agent-relative view of the ground of prohibition on people harming people has been defended by Thomas Nagel in some of his work.[26] It might also be said that there can be *reasons* to do one thing rather than another only for persons because they can have conscious appreciation of considerations for and against acting in some way, but there are no such "reasons for" machines if they have no conscious appreciation of considerations for and against acting in some way. Another ground for thinking that moral principles that apply to people do not apply to machines is that human persons would have emotional responses to and struggle emotionally with killing in some ways, and this cost to them should be taken into account in determining what they should do. But machines do not have emotional responses or struggles that should be taken into account in determining what they should do.

Consider objections to these three views in turn. Suppose that the ground of the wrongness of acting in Topple and Bomb is not concern for the agent's personal view of acting in certain ways (or even for the relationships between agent and victim generated by acting in certain ways). Rather suppose the wrongness is grounded in concern for the potential victim's status as a being who may not be treated in certain ways.[27] Then the potential victim's status could be violated as much by a machine as by a person. It is interesting and important that which theory correctly grounds nonconsequentialist prohibitions on harming could be relevant to which principles to use in programming machines. For purposes of this discussion I shall assume that the "victim-focused" account of prohibitions is correct.

Second, reasons for an entity's behaving in a certain way can exist independently of an entity's awareness of this. Certainly there can be a reason for a person not to drink a poisoned liquid though he isn't aware it is poisoned. Could there similarly be a reason for a machine not to kill in Topple if a person's status provides grounds for his not being killed in that way? We could at least say that there is a reason for there not to be entities that would kill someone in this way and people who were aware of these reasons might be morally required to interfere with the machine that would kill in this way. This could be true even if the machine was not designed by people but fell like manna from heaven or grew like a plant.

The fact that machines would not be affected emotionally by their harming others is irrelevant to the permissibility of their movements if the impermissibility of people behaving in comparable ways has nothing to do with the emotional costs to them of doing so. This would be so if the reason for not behaving in that way stems from the status of the potential victim and the emotional costs arise from implicit recognition by agents that they have acted impermissibly. If people could take a pill that made them not react emotionally to their killing in Topple, this would not affect the impermissibility of their so acting. (The absence of emotional effects, however, might make it easier for machines than people to do the right thing when the cost to people of doing so would be great, e.g., their own destruction. Machines would not have excuses that people might have for not doing the right thing when this involves damage or harm to them. I shall return to this point later.)

Finally, at least some self-driving cars would be programmed in advance to deal with any upcoming situation while a person driving a car would decide what to do when in the situation. Does this make what the car should do different from what a person should do in the same situation? In the Mammography Case it was said that an ex ante decision to use a diagnostic test in a population could be morally permissible even though we know that it will unavoidably harm someone at a time when we will be unable to help her. But that does not mean that we should

use such a mammogram test instead of one that could detect and interfere with harm it was about to cause. Similarly, that a car will be programmed in advance (unlike a human person) does not mean that it should not be programmed to behave in that situation to avoid harming someone in the way a person should. We need not program a device in advance to topple someone in the Topple Case because minimizing deaths would be ex ante in the interests of all if it would be wrong on victim-focused grounds for a person to commit in advance to doing the same thing at a time when she could still avoid doing it.[28]

2.5.3. Programmers and Company Duties

2.5.3.1. Programmers as Bystanders?

Do those who program cars they will not drive occupy a role analogous to that of the bystander in the Bystander Case? One difference between these two agents is that the bystander is dealing with a trolley that is already doubly out of the driver's control: the driver cannot control it as it heads to the five, and he lacks the ability to turn it away from the five to the sidetrack. By contrast, the programmer, at least in the Partial Case, is deciding whether to make the car be to some degree beyond a conscious driver's control. The program is designed not merely to recommend a course of action to the driver but to actually compel the car to make certain movements. (In deciding whether to make completely driverless cars, programmers are deciding whether there is to be any person driving at all.)

Prima facie, the Partial Case seems to raise special moral issues not raised by the Complete Case since it involves deliberately limiting the liberty of conscious agents to decide for themselves whether and how to prevent harm they would cause to others. The bystander in the following revised Bystander Case seems more like the programmer in a Partial Case: A bystander sees a trolley that can be stopped from killing five people only if it is turned where it will kill one other person. The driver retains the power to redirect the trolley, but she may or may not actually do this. The bystander presses a switch that takes the power to turn out of the driver's hands and puts it in his own hands. Call this the Intrusive Bystander Case. It raises at least two questions: (1) Is it permissible for a private person like the bystander to transfer power to himself? (2) Should a bystander as willingly take it upon himself to redirect the trolley, thereby killing a person, when it is possible that the driver herself would fulfill her responsibility to redirect if she retained power? (As we shall see, the answer to questions analogous to these is complicated in Partial Cases by the fact that, unlike what is true in the Trolley and Bystander Cases, the driver herself might be harmed as a result of a decision. Until further notice, I will assume that the driven cannot be harmed.)

The program that takes over in the Partial Case and those who create the program seem like intrusive bystanders. They become less intrusive if the driver has the option of turning the intrusive program on or off when using the car. The latter variant seems most analogous to a variant of the original Trolley Case in which the driver himself relinquishes the power to make a decision in a dangerous situation and hands this power to a bystander.

Ordinarily, it is only governments or their agents that are permitted to act like intrusive bystanders. And ordinarily when government takes on the role of a fully intrusive bystander it is because it has some duty, not a mere permission, to protect people (e.g., to prevent citizens from harming other citizens). This contrasts with the bystander in the Bystander Case, who is thought to have only a permission rather than a duty to redirect the trolley even when the driver has no power to act.[29] As a private person, perhaps the bystander has no right to deliberately take still-retained power away from a driver without the driver's consent, especially if it is power over his own trolley. (Notice that this is consistent with the bystander permissibly interfering in some other way with the driver's controlling the outcome. For example, suppose the bystander quickly pushes a boulder on the track so that it stops the trolley from hitting the five. When the trolley hits the boulder the trolley is also diverted toward killing another person. This side effect need not make it impermissible for the bystander to act.) On the other hand, since the driver is a threat to five people, why is he not liable to having his power over his vehicle being deliberately removed when this cannot harm him and will ensure that appropriate diversion of the threat takes place?

2.5.3.2. Company Agents

Are there additional reasons why programmers have a right and also should be willing to do what intrusive bystanders may possibly not have a right to do or perhaps should not be willing to do? Arguably programmers have such a right and should be willing to act on it because they are agents of a company that is in a distinctive position of producing cars that may cause harm. On one view this distinctive position implies that if car producers can make cars that in morally permissible ways cause fewer casualties, other things equal, they have a duty to the community to do so. This contrasts with a second view, that the company has duties only to the purchasers of its cars and to its stockholders.[30] It also contrasts with a third view, that the company has none of these duties and is at liberty to make any sort of car it wants. If potential drivers do not like it, they can refuse to buy it, and if the company wants to stay in business they will have to change their product.[31] In what follows, I will consider only what the first view may imply.

Unlike the bystander in the Intrusive Bystander Case, programmers are employed by producers who are (in part) analogous to those who made the trolley that malfunctioned and endangered people in the original Trolley Case. Indeed,

for them the choice may be closer to killing more or killing fewer people rather than letting more die or killing fewer (as it is for the intrusive bystander). Some may say "Cars don't kill people, people do," and indeed in Partial Cases it could be the driver's failure that causes a problem. But even in the latter case, a company might have a duty and also reasonably want their product, other things equal, to cause fewer rather than more deaths at the hands of the driver (e.g., by creating cars that will not start until drivers satisfy an in-car device that tests for alcohol level).

In addition, the company programmer is not intruding after the car has been purchased but ex ante. So he need not be changing what the buyer could expect at the time of purchase. This is also unlike the bystander in the Intrusive Bystander Case who would first get involved at the time that redirection is needed. (This difference would be present even if we imagined that the intrusive bystander was the producer of the defective trolley who had not acted ex ante to reduce deaths.)

If the company is determining what a car that might threaten people through the car's failure alone should be programmed to do, it seems appropriate that they think of themselves as programming *requirements* on how the car should move. This would be comparable to (what many consider) an obligation (not mere permission) of the trolley driver to move the trolley so as to minimize those killed (at least when he would not be harmed). This contrasts with merely providing the car with driver-initiated options for minimizing those killed. Programming requirements seem clearly called for in the Complete Case, where company programmers are deciding what a car that on its own will kill some should be programmed to do so as to kill the fewest in a permissible way.

2.5.3.3. Programmers and Drivers

So far we have considered the role of programmers solely in relation to one possible duty of their company regarding its product: to reduce numbers of people killed. Now consider whether and how this duty may combine with some duties specific to drivers. If the original Trolley Case is a guide, then in the Partial Case the driver would have a duty, other things equal, to redirect so as to minimize deaths, at least when the driver is not at risk of harm. Then programmers who are agents of the company might be in a stronger position to program a car ex ante so as to bring about a death-minimizing outcome because that outcome corresponds to one that would result from a driver's doing his duty. However, there could be conflicts between the driver's duties and the company's duties. For example, suppose that if the driver diverts from killing five, the one he will kill is his child. Presumably he does not have a duty to divert and may even have a duty not to divert. A company program to reduce deaths would divert the car. Even though the driver would not actually be responsible for killing his child if the program diverted the car, it seems wrong for the company to ignore a driver's

moral permission or duty not to divert even when it is the car that malfunctions. Perhaps cars could be programmed to act on such personal information from a driver once a lie-detector device in the car had passed it as reliable. (I will discuss the relation between company's and drivers' permissions and duties further later.)[32]

2.5.3.4. Programmers as Drivers

Another difference between company programmers and the bystander is that the latter is assumed not to be involved as either a driver or as a potential victim. But ex ante those who program cars can also reasonably suppose that ex post, when the car will do as it is programmed, they might be one of those involved in a situation in which people are threatened (e.g., either as the driver or one of those outside the car threatened by it). So unlike the bystander, in the Partial Case they could be helping themselves to bring about an outcome they would have a duty to bring about as a driver. They also stand to benefit (or be harmed) from programming decisions they made because the lives saved (or taken) may be their own. This could affect the prudential rationality for them of programming in a morally permissible way even if it does not result in a moral obligation to program in that way.

2.5.4. Pedestrian Liability

A fourth issue to consider is that in the standard Trolley Cases all the people who might die are thought to be equally innocent, nonthreatening individuals whose actions do not make them deserving of or liable in virtue of their actions to being killed by the trolley. (This is on the continuing assumption that the driver cannot be harmed by any action that helps others.) But suppose five people irresponsibly run in front of a car against a red light. (Call this the Irresponsible Five Case.) This does not make them deserve to be killed, but it might make them liable to being killed rather than one innocent person who would be killed if the car were redirected away from them. Desert and liability are commonly distinguished in the following ways: Giving people what they deserve even if it is something bad is thought to be intrinsically good if it is proportional to what they have done. Doing something bad to someone because his actions have made him liable to have it done is consistent with the bad being out of proportion to what he has done, with it being regrettable that the bad thing must be done to him to prevent harm to someone else not liable to bear it, and preferable that there be another way to achieve this good end.[33]

This Irresponsible Five Case could be one in which a car program should not minimize the number of people killed because not everything else is equal. This is because the degree of liability is a morally relevant difference between the five people and the one other person who would be struck in diversion, making everything not be equal among them.

There might even be a case in which several people are hurled at a stationary car (e.g., by a tornado) and would be killed by impact if the car were not redirected. In this case if they impact the car, harm would come only to them and to no one else. However, it may still be best to treat them as innocent threats because their trajectory, for which they are not responsible, causes a problem that could be avoided only if the driver diverted the car, thus killing a pedestrian. I suggest that if they could divert themselves at some moderate cost to themselves to prevent the pedestrian's dying through diversion, they should do so. Furthermore, it seems that their responsibility to do this is greater than that of a mere bystander to pay the same cost if this would prevent the death of the pedestrian. This is so even though they and this bystander are equally morally innocent of causing the problem situation. People may simply have a duty at some moderate cost to correct the inappropriate location of their body.[34] If they are unable to do so, their having this duty may make it permissible for others to impose at least the same cost on them. It might also be argued that it is permissible for others to allow them to be killed by impacting the car rather than have a driver redirect to a nonthreatening pedestrian who would be killed. This is so if costs they could permissibly be made to bear exceed those that are grounded in their personal duties to make moderate sacrifices.[35]

If these claims are correct, then a company would have a duty in programming to take account of liability to be harmed and moral susceptibility to have harms imposed in addition to any duty to minimize lives lost. (For simplicity, I will here. include both these under an extended notion of liability to be harmed.) Liability might either constrain reducing numbers killed to some degree or possibly have lexical priority over reducing numbers killed. Hence it would be important for a programmed car to be able to detect not only the number of people whose lives are at stake but their degree of liability to be killed. There may be heuristics for detecting this. For example, a car could detect if a pedestrian was crossing against a light or if a driver was speeding. Possibly, probability of liability based on evidence of past differential liability in different circumstances could be used in a program.[36] If cars could not be programmed to detect relative liability, would it be correct to program them at all? Possibly it would be if cases requiring ability to detect liability were rare enough, if drivers were no better than programmed cars at determining relative liability, or if drivers were no more likely than a programmed car to behave on a correct determination of liability.

2.5.5. Driver Liability

2.5.5.1. Principles

As noted earlier, in the standard Trolley Cases neither the driver nor any passengers on the trolley are at risk of being killed. So far in discussing moral issues in Partial and Complete Cases I have been assuming this is so as well. Let us now drop the assumption that the driver cannot be killed instead of some pedestrians and consider his liability to be harmed. As noted in section 2.3, I consider these cases to be Innocent Threat rather than Trolley Problem Cases.

In the course of her 2008 discussion of the Trolley Problem, Thomson claimed that if a nonculpable driver of a car will kill nonthreatening innocent pedestrians unless he redirects, he has a duty to do so even if he is the one who will then be killed.[37] If he has this duty, it is not because he deserves to die or is even at fault. Nor is it simply because sacrificing himself will decrease the number of people killed (if it would), for individuals may have a morally sanctioned personal prerogative not to sacrifice themselves for that goal even if social institutions had to pursue it. Thomson does not explicitly say but she may think the driver's duty (which supersedes his prerogative) arises from his being responsible for setting in motion a car that can kill innocent people.

But note that if the driver has this duty to divert at the cost of his own life because he started the car, then he could have this duty even when only one other person (not a greater number) would be killed by him. Indeed suppose it were possible for two drivers to be responsible for driving the car (or for a passenger giving directions to a driver being jointly responsible with the driver for the car's movements). Then the two of them could have a duty to redirect even if this kills them rather than one innocent pedestrian who would otherwise be hit by their car. So an argument for a driver having a duty to impose the death on himself need not be based on and could conflict with reducing the total number of people killed.

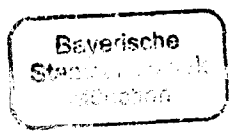
Some may find Thomson's conclusion in her driver case hard to accept because the driver himself has to do what will kill him. A driver who did not do this would most probably be morally and legally excused. However, excuse is not the same as justification, so he may still have failed to do what was right.[38] Furthermore, it would not be as hard for a bystander or a programmer to do what imposes the loss on the driver by either redirecting the car or programming it to redirect, and so the grounds for the driver's excuse would be eliminated. In addition the driver might be liable to have this done to him because he started the car even if he has no duty on these grounds to divert himself and so is even justified in not diverting himself.

Others may reject these conclusions because they think that innocent pedestrians share the liability to be harmed since they are as causally responsible for an accident as the driver simply by being where the car is going; if they were not there, there would have been no accident. I do not accept this view for it seems that a crucial difference between a driver and a pedestrian is that pedestrians are not entities that can damage or harm others on impact in the way cars can harm pedestrians.[39] I acknowledge that this is a complicated issue and moral conclusions could change depending on how this issue is decided.

Some also think that the driver's liability could depend on her reasons for driving. For example, driving an ambulance to fulfill a duty to save other lives might so strongly justify imposing ordinary risk on others that it reduces a driver's asymmetrical liability to suffer any harm relative to pedestrians.[40] But soldiers may have a duty to fight, and yet it is commonly thought they rather than noncombatants (even of the enemy country), who could be considered analogous to pedestrians, should absorb harms in war. It is not my aim in this paper to settle who among the nonrisky, nonnegligent individuals is liable and on what ground. I am primarily concerned to emphasize that wherever it is determined by argument that liability for bearing harm should lie when someone must be harmed, companies could have a duty to take that liability into account in designing programs. This could conflict with other prima facie duties they may have to minimize those killed and to protect the driver.

This raises the practical issue that people may seek to buy cars that are programmed to always favor survival of the driver. (This is so even though when they think of themselves as possible pedestrians on other occasions, they might not favor such a program. In a televised discussion with a philosopher about self-driving cars, a Public Broadcasting Service interviewer, thinking of herself as a driver, said that if there is an accident, she wants to be the one to survive. The philosopher interviewed did not respond that this may sometimes not be the morally correct outcome.) Responding to consumer demand, companies might seek to minimize lives lost and take into account liability to bear harm so long as this applied to everyone besides drivers. Companies could decline to program so as to achieve the morally correct outcome in the light of drivers' duties or liability to bear harm to any greater degree than drivers would ordinarily do so on their own.

If one wanted companies to do more than this, one could try to get them to show that all people would come out better if all drivers were held to programs with higher standards (as in solutions to Prisoners' Dilemmas). Alternatively, one might insist on government regulation to ensure moral solutions that take into account drivers' liability to being harmed and to ensure uniformity in the programs installed. No producers should get a business advantage by providing an "immoral" car that does not incorporate at least "minimal morality" regarding



numbers of lives saved constrained by appropriate liability considerations simply because this will increase their sales.[41] On the other hand, producers and governments should not require the production of what would also be immoral cars, imposing burdens on drivers to which they are not liable and to which they do not consent, for the sake of always maximizing lives saved despite the liability of pedestrians. However, producers should not necessarily be prohibited from programming at the request of drivers altruistic (or supererogatory) cars that allow drivers to bear burdens beyond both those they owe (which a "strictly dutiful car" could ensure) or which may permissibly be imposed on them.

Finally, note that it is easier for a driver to *buy* a car that she knows is programmed to do what will sacrifice her than to actually sacrifice herself; buying such a car is not a way to sacrifice oneself. This is in part because it is uncertain whether one will ever be in a situation where one would be sacrificed by one's programmed car. Combined with an ex ante desire to do what might be one's duty in a tragic case and foreseeing one's lack of courage to do it at high cost, some drivers may actually want to buy a car that is programmed to produce outcomes that track either their duty or liability to bear harms (where these differ). (This is in addition to such drivers considering the possibility that they will at times be pedestrians.)

2.5.5.2. Illustrative Cases

To reinforce these conclusions consider some cases.

2.5.5.2.1. Case 1

Suppose (for the sake of argument) an in-control driver would have a duty to sacrifice himself by diverting rather than kill innocent nonthreatening pedestrians. Does this imply that company programmers have permission or even an obligation to program a threatening car in the Partial Case to divert from killing more nonthreatening pedestrians to killing fewer driver(s) of the threatening car? It would not be surprising if the driver couldn't be trusted when he is in control of the car to do his (assumed) duty to minimize lives lost when doing so would cost him his life. So without forcing anyone to sacrifice himself, they would arrange for the car to generate an outcome that corresponds to that from the performance of a duty that (arguably) *both* the producer and the driver have in this case to minimize lives lost even at the driver's expense.

Suppose the driver had no duty to sacrifice himself. He still might be liable to have costs imposed on him by others. Suppose he was no more liable in virtue of driving to bear costs than anyone else. In diverting the car when this kills him the programmers would still treat him no worse than they (or he) would treat an innocent pedestrian in diverting to her to minimize innocent pedestrians killed,

though she has no duty to sacrifice herself and is not liable to bear harm in virtue of her actions.

2.5.5.2.2. Case 2

Suppose a driver would have a duty to divert at the cost of his own life to save pedestrians. Does this imply that the company should program to divert a car from killing fewer (or the same number of) nonthreatening pedestrians in a way that results in the death of more (or the same number of) driver(s) of the threatening car? If programmers did this, they would not act on any producer's duty to minimize lives lost due to its machine. Their disfavoring the driver(s) would have to involve either arranging for an outcome because it corresponds to one that would result if a driver performed his duty to sacrifice himself and/or involves imposing harm to which the driver is liable even in the absence of his duty. Acting on this consideration would override the producer's other prima facie duty to minimize lives lost since more lives might then be lost. If a company has a duty to take account of liability to bear costs in programming cars, then while it may be contentious that the driver is liable if he is, the company has its own duty to take account of his liability even if this does not minimize lives lost.

2.5.5.2.3. Case 3

Should the company program to *prevent* the diversion of a threatening car from killing both one innocent nonthreatening pedestrian and the driver of the car toward killing two innocent nonthreatening pedestrians? In this case the driver would lose his life not in being diverted but in not being diverted; his life would be saved by diverting toward more pedestrians. The same number of people would be killed either way.

This case raises the following question: When costs to a driver would be high, could there be a moral difference between (1) a program preventing a driver from increasing the number of pedestrians killed and (2) a program reducing the number of pedestrians a driver kills? Suppose there is such a moral difference, in favor of (1). Then it may be permissible for a program to at least prevent the driver's life being saved by diverting when more pedestrians will be killed (though the same number of people would be killed). This could be so even if the program should not lethally divert the driver in order to prevent his killing these pedestrians. Hence if he were headed to killing two people but would survive this, such a program would not divert him toward one other pedestrian when he would die. However, the program would prevent his diverting from killing one pedestrian to killing two, though he will not survive without the diversion. (If the driver's life had the same moral weight as a pedestrian's, preventing his diversion in the latter case would yield the same outcome as not turning from killing one set of two people toward killing another set of two people.)

However, suppose the driver would kill two pedestrians and himself if the car is not diverted, and he would be saved but kill two other people if the car is diverted. In this case, if the car is diverted, an additional life (of the driver) would be saved and there would be no increase in the number of pedestrians killed. Even if there is something to be said against killing two other people rather than letting those originally threatened be killed, saving the life of a driver who does not deserve to die seems important enough to justify diversion. (If the driver himself were in control, he could permissibly save his life in this way since he wouldn't be increasing the number of pedestrians killed.)

2.5.5.2.4. Case 4

Should the company program to divert a car from killing more nonthreatening pedestrians in a way that kills the single driver of the car rather than in an alternative way that kills one different nonthreatening pedestrian? In this case, the car producer's duty to minimize numbers killed would be satisfied either way. If the driver is liable to bear harm, the producer should arrange for the car to divert in a way that kills the driver rather than in a way that kills a different pedestrian.

The overriding conclusion of the discussion of cases 1–4 is that even if producers should not bring about outcomes simply because they correspond to the performance of drivers' duties, bringing about the same outcome or even one more burdensome for the driver can sometimes be necessary in order for producers to carry out their own duty to take account of a driver's liability to bear harm.

2.6. Passengers and Other Drivers

In the two standard Trolley Cases (with a driver or a bystander called on to redirect), the five potential victims and the one to whom the trolley could be redirected are not imagined to be either on the threatening trolley or on another trolley. But programmers for cars recognize that a car can face a collision with other cars and that people initially and potentially threatened might also be in the cars. Rampolla and MIT's Moral Machine website present cases in which if redirection occurs, nondrivers who would be harmed are passengers in the threatening vehicle. I do not think they deal with cases in which those initially threatened are also in vehicles or those potentially threatened are in vehicles other than the initially threatening vehicle. Let us consider a variety of such cases.

2.6.1. Other Drivers

Suppose for the sake of argument that a driver is liable to be sacrificed relative to nonthreatening pedestrians she would otherwise hit simply because she is

driving a threatening car. Are drivers in nonthreatening cars also liable to have harm due to a threatening car redirected to them rather than to nonthreatening pedestrians? For example, suppose the driver in the car threatening pedestrians cannot prevent harm to them by herself bearing the cost. She or the program running her car can either redirect toward another car with a driver or to another pedestrian. Choosing the car with the driver might be justified by analogizing the case to players in a dangerous game (in this case, car driving) who should when possible confine themselves to injuring one another rather than nonthreatening nonplayers if someone must be injured. If so, cars should be programmed to detect and at least sometimes redirect to other cars rather than to pedestrians even when those cars have drivers and diverting in this way will not minimize deaths. "Playing the game of driving" would then be another source of liability to harm.

When a threatening driver could bear costs to prevent harm to others, additional issues related to liability to bear costs will arise: Should the number of all drivers' lives at stake in a decision, possible fault, or merely who is the initial threatening driver determine programming? Could the programmed car detect and "act" on those factors at least as well as unassisted drivers?

Perhaps in some cases involving multiple vehicles it may be possible to distinguish between something like an offensive and a defensive threat. An "offensive" innocent threat would be presented by the Partial or Complete car that initially nonintentionally and nonnegligently threatens either pedestrians or other drivers. A defensive threat might be presented by a vehicle that has to respond to that initial threat; doing so may result in its threatening either the initial threat, other vehicles, or nonthreatening pedestrians. Even if an offensive threat should be programmed so that its driver is sacrificed rather than another driver to whom he presents a threat, the driver who becomes a defensive threat may be liable to be harmed rather than pedestrians his vehicle might harm. (This is because of his participation in the dangerous practice of driving.)

2.6.2. Passengers

Aside from the drivers of threatening vehicles, and even in Complete Cases where drivers are absent, there may be passengers in vehicles.[42] Suppose nondrivers in the threatening vehicle would be killed if the car were redirected from harming nonthreatening pedestrians. Should vehicles that can detect the presence of people inside the car be programmed to count their lives on a par with pedestrians threatened by the vehicle and do what reduces the number of people killed?

People often voluntarily decide to be passengers in a vehicle that they know potentially threatens pedestrians. Furthermore, the vehicle might not have

started at all in the absence of passengers, and passengers may tell vehicles where they want to go if not how to get there. For example, one of the benefits of completely self-driving cars is that they would increase the mobility of blind and paralyzed people. But if such passengers gave a command for the car to start, are they not like the driver who started the trolley that then went beyond his control (even if the passengers never drove the car to begin with)? Suppose such drivers should be given less weight than pedestrians because they are liable to bear costs. Then shouldn't passengers whose directions start a car also be given less weight than pedestrians by a program for Complete Cars?

What about voluntary passengers who did not start the car but chose to join those who did start the car? Their joining a dangerous game provides a ground for some liability to bear costs.[43] When a car with seven passengers threatens two pedestrians perhaps, each of the seven should count for only a fraction of a person in a calculation of lives lost. However, joining oneself to a car before it becomes a threat is still different from, for example, hopping a ride on what one knows is a vehicle headed to killing pedestrians. In the latter case, the passengers' lives should certainly have reduced weight relative to the lives of pedestrians when deciding whether to divert the vehicle for they knowingly attach themselves to a threat that should have been diverted. If possible, a car should be programmed to detect such morally relevant differences among passengers.

2.6.3. Mere Cars

Do the following conclusions at least seem certain? When a car with a driver and/or passengers threatens a completely empty car, the empty car should be destroyed rather than kill pedestrians or other drivers and passengers. Also, empty Complete Cars should be programmed to "sacrifice themselves" rather than pedestrians or drivers and passengers in other cars even when the latter are causally or (sometimes) even morally responsible for the initial problem. These conclusions seem to follow from the view that lives of persons take precedence over property, at least when there is no intentional wrongdoing by those people that aims at destruction of property.

However, what if the empty Complete Car is the only one that can be sent to save many other people? I think it (like the Ambulance) cannot be allowed to run over even fewer nonthreatening pedestrians in order to avoid its own destruction and continue on its mission. But are those who are morally responsible for being in harm's way (like those in the Irresponsible Five Case) liable to being harmed by the car rather than having it destroyed when it is necessary to save many other lives? Being liable to bear costs rather than have the car redirected to *kill* others,

as is true of the five in the Irresponsible Five Case, is still morally different from being liable to bear costs so that the car can go on to *save* others, I think.[44]

Notes

1. F.M. Kamm, *Morality, Mortality*, vol. 2 (New York: Oxford University Press, 1996).
2. I am assuming that the Partial Cases involve taking control away from a driver who is not allowed to drive only in certain dangerous situations. Other types of Partial Cases may involve self-driving cars that turn over control to the driver only in certain dangerous situations. In addition, it is said that self-learning cars need not be programmed. For simplicity's sake, I will speak of cars that decide what to do as acting on a program.
3. Though the case was created by Philippa Foot in "The Problem of Abortion and the Doctrine of Double Effect," in *Virtues and Vices* (Los Angeles: UCLA Press, 1978): 19-33, it was only later called the Trolley Problem in Judith Thomson, "Killing, Letting Die, and the Trolley Problem," *Monist* 59, no. 2 (1976): 204–17.
4. See Judith Thomson, "The Trolley Problem," *Yale Law Journal* 94 (1985): 1395-1415. In a switch from her 1976 article, Thomson came to apply the term "Trolley Problem" to this case alone, though others did not.
5. In another type of case, which I call Crosspoint, a bystander must decide whether to turn a trolley that will kill many if it remains at a crossing point toward killing five or killing one. Here the bystander is choosing between letting many die or killing either five or one.
6. Such a case was also introduced by Thomson, who called it the Fat Man Case in her "Killing, Letting Die, and the Trolley Problem."
7. See Chris Rampolla, "The Trolley Problem Reimagined: Self-Driving Cars," *Aero*, March 31, 2017. I was directedto this site by a nonphilosopher professor who works on the Trolley Problem and who saw it as popular discussion that provided a good introduction to the problem for the general public.
8. Rampolla actually speaks of the car, not the driver, striking the pedestrian and the driver possibly causing a collision that saves the pedestrian but will likely kill him. Reserving the term "killing" only for the latter effect seems a biased description.
9. The problem of applying results achieved by assuming certainty when real life presents us only with risks is a topic discussed by others. See note 25.
10. Neither a nonphilosopher professor nor a postdoctoral psychologist working on the Trolley Problem, both of whom were present when I discussed Rampolla on April 23, 2018, had ever heard of Innocent Threat Cases. Perhaps they would still have realized that different moral problems might be raised by them from those raised by the standard trolley cases. However, Joshua Greene, a philosopher and psychologist who has written about the Trolley Problem, gave as a real-life example of it (when speaking at a Safra Ethics Center conference dinner at Harvard in 2017) a case in which doctors must decide whether to confine a person carrying a contagious disease, thus

imposing costs on him in order to save others from the disease. He reported that doctors were concerned about imposing costs on one person to save others, and he gave the impression that this was like the concern about turning the trolley on one person to save others. But in the medical case costs would be imposed on the person who presents the threat and doctors should not, I think, be as concerned about imposing a cost on an innocent threat as on an innocent nonthreatening person.

11. See Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974).
12. See Saba Bazargan, "Killing Minimally Responsible Threats," *Ethics* 125, no. 1 (2014).
13. Though killing the driver might stop the threat, as when killing him causes his body to fall on a brake that is otherwise inaccessible to him.
14. Prof. Bert Huang refers to a psychological study on trolley problems in which researchers expose subjects to "the actual argument—that no one is obliged to sacrifice his own life to save others, and that it seems immoral to force another to make a sacrifice one would not have to make oneself." B. Huang, "Law and Moral Dilemmas," *Harvard Law Review* 130 (2016): 673. This description suggests that the person who would die from the trolley if it is redirected toward him in the standard cases is made to sacrifice his life. But this ignores the possibility that there is a conceptual (and moral) difference between imposing loss of one's life on oneself (either voluntarily or coerced) and having that loss imposed on one by another.
15. In sum, the following are among the cases we have so far distinguished: (1) someone who threatens but cannot be threatened decides on which nonthreateners to impose losses (the standard driver Trolley Case); (2) a bystander who cannot be threatened decides whether to impose losses on nonthreateners with no possibility of harming the threatener (the Bystander Case); (3) someone who threatens and can be threatened decides whether to impose losses on himself or his potential nonthreatening victims (an Innocent Threat Case as I use the term); (4) a bystander who cannot be threatened decides whether to impose losses on a threatener or allow (or cause) losses to be imposed on nonthreatening victims (another Innocent Threat Case).
16. This is also true in many cases presented by the MIT Moral Machine website, http://moralmachine.mit.edu.
17. For objections to Foot's proposal, see F.M. Kamm, *The Trolley Problem Mysteries* (New York: Oxford University Press, 2015).
18. Marya Zilberberg, "Medicine as the Trolley Problem," *Healthcare, etc.*, July 25, 2012, http://evimedgroup.blogspot.com/2012/07/medicine-as-trolley-problem.html.
19. Bioethicist Nir Eyal once suggested this (much before mounting evidence about the dangers of e-cigarettes). The example used merely as a hypothetical case remains useful for my purposes.
20. The Ambulance Case with which I began this discussion also contrasts with diverting the trolley. Though it involves choosing between saving five (by getting them to the hospital) and killing one, it is not like the Bystander Case (which also involves choosing between saving five and killing one). It may seem that one way of marking the distinction is that, unlike the Trolley, the Ambulance (like the Gas or Bomb) that would kill the one is a new threat that is not already threatening the five. But the case (presented in the text) in which the diverted trolley causes a new rockslide that kills

one person shows why it is not quite right to say that the problem in the Ambulance Case is that a new threat kills the one. It would be better to say the following: It is not to remove a threat that the ambulance itself presents to the five that we would consider having the ambulance continue on killing one other person, and so it is impermissible to drive the ambulance over the person on the road. However, elsewhere I have considered what I call the Lazy Susan Case to be like the Trolley Case, though it does not even involve redirecting a threat from the five. Rather it involves moving people away from a threatening trolley that cannot be redirected. In one version, their being moved will create a new rockslide that kills a bystander. I have argued that doing what saves the five but kills the one is as permissible in this case as in the standard trolley cases, even though the one person is killed neither by what threatens the five nor by a new threat created by the threat itself moving away from the five. I argued that the Lazy Susan Case is like the Trolley Case because the same principle explains the permissibility of both turning the trolley and turning the Lazy Susan. In my view, this principle (put roughly and far too simply) is that what kills one other person just is the five being saved. The trolley turning away which kills the one does not merely cause the five to be saved (as the bomb in the Bomb Case would); it constitutes the five being saved. Similarly the five being moved away from the trolley on the Lazy Susan, which leads to one other person dying, just is the five being saved. For further discussion of this, see F.M. Kamm, *Intricate Ethics* (New York: Oxford University Press, 2007) and *The Trolley Problem Mysteries* . I omit further discussion of Lazy Susan–type cases here since they do not seem pertinent to real-life cases of self-driving cars.

21. I do not in this chapter defend the correctness of these judgments and the principles that justify them (though I said something about this in note 21). Thomson herself came to reject her earlier judgment that it is permissible for the bystander to redirect the trolley, thus killing one and saving five. See Judith Thomson, "Turning the Trolley," *Philosophy & Public Affairs* 36 (2008), 359-374 and my discussion of her later view in *The Trolley Problem Mysteries*.
22. Principles and rules are commonly distinguished. For example, H. L. A. Hart conceived of the legal system as consisting of rules. By contrast, Ronald Dworkin thought that system fundamentally consisted of principles (such as "not benefiting from one's crime" or "fair play") that required more interpretation than rules and that could ground rules. The principles might also guide us when rules run out or when rules lead to conclusions in particular cases that are inconsistent with grounding principles.
23. Fiona Woollard and Will McNeill discuss these issues in their "Driverless Cars" and "Ethics without Algorithms," both unpublished manuscripts.
24. I continue to put to one side the important issues of certainty versus probability of deaths and knowledge of this. In the standard hypothetical trolley cases one assumes certain death for the five if the trolley continues or for someone else if the five are saved. One also assumes knowledge of this by the decision-maker. This is not necessarily true in real life. On this problem of applying Trolley Problem reasoning to programming cars, see Sven Nyholm and Jilles Smids, "The Ethics of

Accident-Algorithms for Self Driving Cars: An Applied Trolley Problem?," *Ethical Theory and Practice* 19, no. 5 (November 2016): 1275–89.

25. Note also that in trolley cases it is because the trolley is no longer under the driver's control in heading to the five (though it is also not self-driving) that a problem initially arises. By contrast, it is because human drivers in cars completely under their control often fail to do the right act that initial life-threatening situations often arise. It is somewhat ironic that (1) cases in which a problem arises in the first instance because a vehicle lacks a driver in control are being looked to (by some) for guidance about what to do (2) when a car's lacking a driver in control is supposed to prevent problems from arising in the first instance.
26. See, for example, Thomas Nagel, *The View from Nowhere* (New York: Oxford University Press, 1986).
27. For defense of a view like this see, for example, my *Morality, Mortality,* vol. 2
28. I am grateful to Jesse Berthold and Arthur Applbaum for raising questions that led to some of my responses in this section.
29. This may be because the bystander will wrong the one person he kills even if he acts permissibly in doing so.
30. Suggested by Larry Temkin.
31. Suggested by Shelly Kagan.
32. We could also consider the role of programmers' duties in relation to possible victims' duties. For example, suppose the five toward whom the car is headed were the parents and guardians of the one person toward whom the car would be diverted. They might have a duty or preference to see to it that the car is not diverted. However, drivers are no more likely to know of such relations between potential victims than programmed cars would, whereas they do know about their own duties and preferences.
33. On the distinction between desert and liability, see Jeff McMahan, *Killing in War* (Oxford: Oxford University Press, 2009). In the Irresponsible Five Case, McMahan would say that the five are liable to the harm because they have "assumed the risk." A crucial issue is whether being liable should be a tiebreaker when all else is equal between potential bearers of a loss or whether it should come with built-in limits on the loss to which one can be liable depending on what makes one more liable than someone else. On the latter view one might be liable to a higher chance of bearing a loss (e.g., 80%) or liable to bearing a loss only up to size x. If someone must bear a certain-to-occur loss or a loss over x, then on this second view who should bear the loss should be determined by giving equal chances. On the first view, being liable wouldn't have such built-in limits and could serve as a tiebreaker that determines on whom the certain-to-occur loss or loss larger than x should be placed. (The number of people harmed might count in determining the size of the loss.)
34. I discussed cases of this sort in F.M. Kamm, "The Insanity Defense, Innocent Threats, and Limited Alternatives," *Criminal Justice Ethics* 6, no. 1 (1987): 61–76.
35. The permissibility of turning the trolley on one innocent person relies on the view that what may be imposed on someone exceeds his duty to impose harm on himself. But if liability depends on a person's action or movement, there would not be this ground for imposing harm on the one person in the trolley case. Furthermore, more

people who are liable to have harm imposed on them might sometimes permissibly be harmed to save fewer people. By contrast turning the trolley and harming the one nonliable person depends on fewer people overall dying.

36. I owe these suggestions for the heuristics to Jeff McMahan, Shelly Kagan, and Larry Temkin.
37. Thomson discusses this case in her "Turning the Trolley," 369.
38. Note that his duty is not necessarily to actively redirect when that would have killed him. For there might be a case in which unless he diverts he will continue on in a way that results in his being killed, but if he diverts to save himself he will kill others. Then he might have a duty to refrain from diverting. (I discuss such a case later in the text.)
39. However, there is the difficulty of distinguishing morally between a moving pedestrian who is walking and an innocent hurled at a car. Why would the latter be liable to bear costs, as I argued earlier, and not the pedestrian since neither directly threatens harm or damage to others?
40. Jeff McMahan holds such a view.
41. Analogously, suppose a college complained that it couldn't attract students if it did not allow some cheating because other schools allowed some cheating and students preferred those schools. The solution is not to give up the correct moral standard but for all schools to agree to enforce the standard. Suppose students would then prefer no education (comparable to people not buying any cars that tracked moral requirements). Suppose this education outcome was bad (as might not be true if as a consequence of not buying cars people used only public transportation). Then it might be necessary to either require the practice of education or make it more attractive in some way other than by allowing some cheating.
42. Unlike the Moral Machine website, I shall consider only person passengers, not non-person animal passengers. That website also considers animal pedestrians.
43. It might be said that passengers (and drivers) stand to benefit from using cars and this is what grounds their liability to be harmed rather than pedestrians. In the case of both passengers and drivers I do not wish to derive liability to harm from standing to benefit. For even if drivers and passengers did not stand to benefit from using cars, the risk of harm to others from the devices they use is an important reason why they might be liable to bear costs when someone must.
44. I am grateful to Mathew Liao for inviting me to write this chapter. I am grateful to him, Shelly Kagan, Jeff McMahan, Larry Temkin, students in my Rutgers philosophy seminars, and audiences at the Edmond J. Safra Ethics Center of Harvard University and at the University of Granada Philosophy Department for comments on earlier versions of this chapter.

References

Bazargan, Saba. "Killing Minimally Responsible Threats." *Ethics* 125, no. 1 (2014): 114-136.

Foot, Philippa. "The Problem of Abortion and the Doctrine of Double Effect." In *Virtues and Vices*. Los Angeles: UCLA Press, 1978: 19-33.

Huang, B. "Law and Moral Dilemmas." *Harvard Law Review* 130 (2016): 659-699.
Kamm, F.M. "The Insanity Defense, Innocent Threats, and Limited Alternatives." *Criminal Justice Ethics* 6, no. 1 (1987): 61–76.
Kamm, F.M. *Intricate Ethics*. New York: Oxford University Press, 2007.
Kamm, F.M. *Morality, Mortality*. Vol. 2. New York: Oxford University Press, 1996.
Kamm, F.M. *The Trolley Problem Mysteries*. New York: Oxford University Press, 2015.
McMahan, Jeff. *Killing in War*. Oxford: Oxford University Press, 2009.
Nagel, Thomas. *The View from Nowhere*. New York: Oxford University Press, 1986.
Nozick, Robert. *Anarchy, State, and Utopia*. New York: Basic Books, 1974.
Nyholm, Sven, and Jilles Smids. "The Ethics of Accident-Algorithms for Self Driving Cars: An Applied Trolley Problem?" *Ethical Theory and Practice* 19, no. 5 (November 2016): 1275–89.
Rampolla, Chris. "The Trolley Problem Reimagined: Self-Driving Cars." *Aero*, March 31, 2017.
Thomson, Judith. "Killing, Letting Die, and the Trolley Problem." *Monist* 59, no. 2 (1976): 204–17.
Thomson, Judith. "The Trolley Problem." *Yale Law Journal* 94 (1985): 1395-1415.
Thomson, Judith. "Turning the Trolley." *Philosophy & Public Affairs* 36 (2008): 359-374.
Woollard, Fiona, and Will McNeill, "Driverless Cars" and "Ethics without Algorithms" (unpublished).
Zilberberg, Marya. "Medicine as the Trolley Problem." *Healthcare, etc.*, July 25, 2012. http://evimedgroup.blogspot.com/2012/07/medicine-as-trolley-problem.html.